Environmental Modelling & Software 82 (2016) 183–217

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Multi-scale land-use disaggregation modelling: Concept and application to EU countries

Matieyendou Lamboni*, [1], Renate Koeble, Adrian Leip

EC-JRC, Institute for Environment and Sustainability, 21027, Ispra, Italy

ARTICLE INFO

Article history:
Received 25 September 2015
Received in revised form
27 April 2016
Accepted 29 April 2016
Available online 9 May 2016

Keywords:
Bayesian modelling
Homogenous spatial unit
Prediction of land-use areas
Spatial disaggregation
Uncertainty

ABSTRACT

Changes of carbon stocks in agricultural soils, emissions of greenhouse gases from agriculture, and the delivery of ecosystem services of agricultural landscapes depend on combinations of land-use, livestock density, farming practices, climate and soil types. Many environmental processes are highly non-linear. If the analysis of the environmental impact is based on data at a relatively coarse-scale (e.g. farm, country, or large administrative regions), conclusions can be misleading. For an accurate assessment of agri-environmental indicators, data of agricultural activities and their dynamics are needed at high spatial resolution. In this paper, we develop and validate a spatial model for predicting the agricultural land-use areas within the homogenous spatial units (HSUs). For the EU-28 countries, we distinguish about 1.5×10^5 HSUs and we consider 30 possible land-uses to match with the classification used in the Common Agricultural Policy Regionalized Impact (CAPRI) model. The comparison of model predictions with independent observations and with a simple rule-based approach at HSU level demonstrates that the predictions are generally accurate in more than 75% of HSUs. The frequent crops or land-use are better predicted. For non-frequent crops and/or crops requiring specific cultivation conditions, the model needs further fine-tuning.

Software

The Land-Use Disaggregation Model (LUDM) aims at predicting the land-use areas within the fine-scale units. LUDM is written in R (R CRAN) and is freely available. The source code is maintained and can be downloaded as a zip file from http://ludm2016.blogspot.com/2016/04/ludm.html. The zip file contains a ReadMe file and accessory files. Use the ReadMe file for suggestions on program instruction and notes on terms of service. LUDM model is free, regulated under the GNU General Public License v3 (http://www.gnu.org/copyleft/gpl.html) and intended for further open-source development.

1. Introduction

The agricultural sector contributes with about 10.6% to the total greenhouse gas (GHG) emissions, excluding Land Use, Land-Use Change and Forestry (LULUCF), in the EU-27 (EEA, 2012) and with 12% to the worldwide emissions (Lehuger et al., 2011). Agri-environmental indicators are used for assessing the impacts of agricultural activities on the environment and they include nitrogen balance, GHG emissions, losses of reactive N from agricultural sources, soil erosion or soil carbon stock changes. Most of the agri-environmental indicators are calculated by combining spatial information on agricultural land-use, livestock density, farming practices, soil characteristics and climate variables at a local, farm, sub-regional or regional scale (Leip et al., 2008; Lamboni et al., 2009; Gocht and Röder, 2014; Follador et al., 2011; Leip, 2011; Leip et al., 2013; Butterbach-Bahl et al., 2011). Reliable environmental impact assessment requires the availability of the data and its dynamics at a high spatial resolution. Many environmental processes are highly non-linear. If the analysis of the environmental impact is based on data at a relatively coarse-scale (e.g. large administrative regions), conclusions can be misleading (Gocht and Röder, 2014). Exceedance of thresholds or the identification of

* Corresponding author. Present address: 228-UMR Espace-Dev, 275 Route de Montabo, 97323, Cayenne Cedex, French Guiana.
E-mail address: matieyendou.lamboni@gmail.com (M. Lamboni).
[1] Present address: University of Guyane, Department DFRST, 97346, Cayenne, French Guiana.

http://dx.doi.org/10.1016/j.envsoft.2016.04.028

hotspots is often more meaningful than average values (Leip et al., 2008; Britz and Leip, 2009; Leip, 2011; Kempen et al., 2005; Gocht and Röder, 2014). For example, nitrate may be leached to the groundwater from intensive green maize cultivations with high fertilizer N input on soils with coarse texture. This effect may be missed if N input is averaged over all cultivations and soils.

The development of highly heterogeneous environmental indicators has been widely investigated, including semantic interpolation and fuzzy similarity approaches; first, second or higher order approximations (Dubois et al., 1997; Mas et al., 2012; Bordogna et al., 2012). Land use modelling is an important approach for evaluating global environmental impact (Pérez-Vega et al., 2012).

When the main factors driving the environmental indicators are highly heterogeneous, quantification of agri-environmental impact faces the challenge of the trade-off between relatively simple first-order methods on one hand and potentially higher accuracy but more cumbersome methodologies on the other hand. We used a first order approach while minimizing the aggregation bias by defining spatial units for which the factors take a single vector of values (homogeneity regarding these factors). Various shapes of fine scale units are conceivable depending on the thematic focus.

Recently, Leip, 2011 created a map of Homogeneous Spatial Units (HSU) for Europe including parts of the Near East and Northern Africa. A HSU is the finer-scale spatial clustering driven by the search for homogeneity with regard to selected environmental factors. It is defined as cluster of km^2 grid cells within a subnational region (e.g. NUTS2/3) which covers an area of similar characteristics in terms of soil, climate and relief. For the current extent of the European Union (EU-28), about 1.5×10^5 HSUs were delineated (excluding evident non-agricultural areas). Corresponding to the chosen grid resolution,[2] the minimum size of a HSU is 1 km^2. The average (resp. median) size of the HSUs in EU-28 is 22 km^2 (resp. 11 km^2). The largest HSUs with areas of up to 566 km^2 occur in zones of high homogeneity of the delineation parameters (e.g. Northern Baltic States).

The land-use areas within the HSUs are required as prior information for developing landscape, biodiversity and other agri-environmental indicators in a series of steps including the disaggregation of animal numbers (a.o. via fodder production) and the disaggregation of the N-input (combining crop requirements, N-availability and local environmental conditions) (Leip et al., 2008; Britz and Leip, 2009; Leip, 2011). They serve also as a link between agro-economic models such as the Common Agricultural Policy Regionalized Impact (CAPRI) model and biophysical models such as framework DNDC-EUROPE (Leip et al., 2008).

Some disaggregation approaches (Reibel and Agrawal, 2007; Kempen et al., 2005; Chakir, 2009; Lamboni et al., 2013; Röder and Gocht, 2013; Gocht and Röder, 2014), mainly the dasymetric mapping approaches, make use of statistical information about the land-use at a coarse-scale, geographically-referenced or remote sensing data and ground-based and point-based observations to predict the land-use areas within the fine-scale units such as HSUs. To be consistent with the statistical data available at a coarse-scale (mainly at an administrative level), these approaches first estimate the prior distribution of the land-use areas at a fine-scale and then constrain the predicted areas with respect to the statistical data available at a coarse-scale using the Bayesian highest posterior density.

In this paper, we propose an extension of the dasymetric mapping approaches by proposing a full Bayesian approach for building the disaggregation model and for predicting the land-use areas inside HSUs. The aim of this paper is to propose a generic Bayesian framework for spatial disaggregation of shares from a coarse-scale (e.g. administrative region) into HSUs. We model the shares inside a given HSU using a multinomial model and we use the disaggregation model for predicting the land-use areas within the HSUs for the all EU-28 countries. The model includes three steps presented in Fig. 1. First, we combine point-based field observations of land-use with the environmental and topographical information, land cover classes and the prices of selected products to get the *a priori* distribution of the model parameters. Second, we propose a transformation that allows for integrating land-use statistics available at an administrative level in order to update the model parameters (*a posteriori* distributions). Third, we propose the constrained prediction of land-use areas inside the HSUs that match with the land-use areas available at an administrative level (NUTS2/3[3]) and with the areas of the HSU units using the quadratic programming (QP) approach. We discuss the results, including the validation of the model with independent high resolution data for France and the comparison of the model predictions with the results of a simple rule-based disaggregation approach.

The paper is organized as follows: we describe the question treated in Section 2 and the data in Section 3. We present the global structure of the model and the estimation of the model parameters in Section 4. We also present the QP problem and the constraints used in the case of land-use areas. Section 5 presents the main choices made in order to get the results by taking into account the particularity of some countries. In Section 6, we discuss the predicted results and we conclude in Section 7.

1.1. Notation

Throughout this paper, we use Fig. 2 to characterize both fine-scale units and coarse-scale units. We suppose that a coarse-scale unit is a collection of fine-scale units that belong to it and a fine-scale unit belongs to only one coarse-scale unit. The fine-scale unit is what we call Homogenous Spatial Unit (HSU). A HSU is one of the possible finer-scale spatial clustering driven by the search for better homogeneity. HSU is a grid cell of 1 km $\times$ 1 km, or a collection of these grid cells having similar properties. A HSU is characterized by its explanatory variables and its area a_h.

Coarse-scale units used are administrative regions at the NUTS2 or NUTS3 level. An administrative region called NUTS2 is a collection of the NUTS3 sub-regions that belong to it. An administrative sub-region called NUTS3 is a collection of the HSU units that belong to it (see Fig. 2). Each HSU belongs to one NUTS3 region. For readability and without loss of generality, we use in the following text the terms “HSU” (resp. “NUTS3” or “NUTS2”) to refer to the fine-scale unit (resp. coarse-scale unit).

Throughout this paper, we use L as the number of all possible and exclusive land-use classes. We use $\mathbf{A}_n = [A_{1,n}, \ldots, A_{l,n}, \ldots, A_{L,n}]^{\mathcal{T}}$ as a vector of land-use areas with $n = 1,2,\ldots,N$. N is the total number of vector of observations of land-use areas inside a given NUTS2 region. As we often have one vector of observations per NUTS3 sub-region, N is the number of NUTS3 that belongs to a given NUTS2. For Germany, we have one observation per NUTS2, e.g. $N = 1$.

Let $\mathbb{P}_r(\cdot)$ denote the probability; $\mathbb{E}(\cdot)$ denote the expectation and $\mathbb{MSE}(\cdot)$ be the mean square error. We use $\mathbb{I}$ as an identity matrix

[2] The grid is based on the recommendations of the 1st European Workshop on Reference Grids in 2003 (Annoni, 2005) and the INSPIRE draft specifications on geographical grid systems (INSPIRE, 2008) with ETRS89 Lambert Azimuthal Equal Area coordinate reference system and the center of the projection at the point 52° N, 10° E and false northing: $Y = 3{,}210{,}000$ m, false easting: $X = 4{,}321{,}000$ m.

[3] Nomenclature of Units for Territorial Statistics, level 2 or 3.

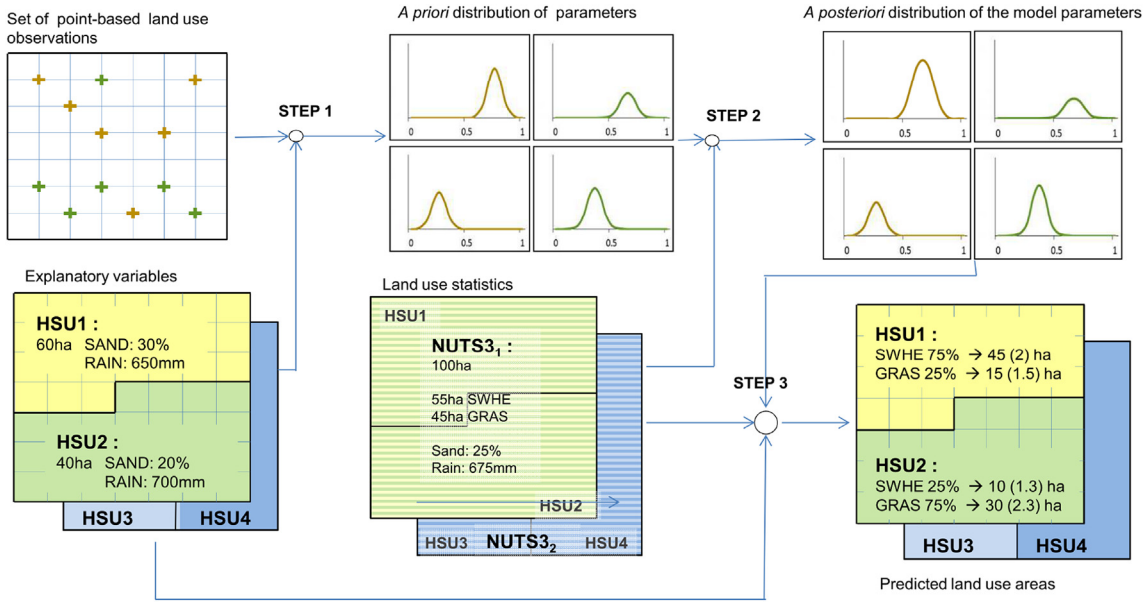

Fig. 1. The three steps of the Land-Use Disaggregation Model (LUDM). For more details see Fig. 4.

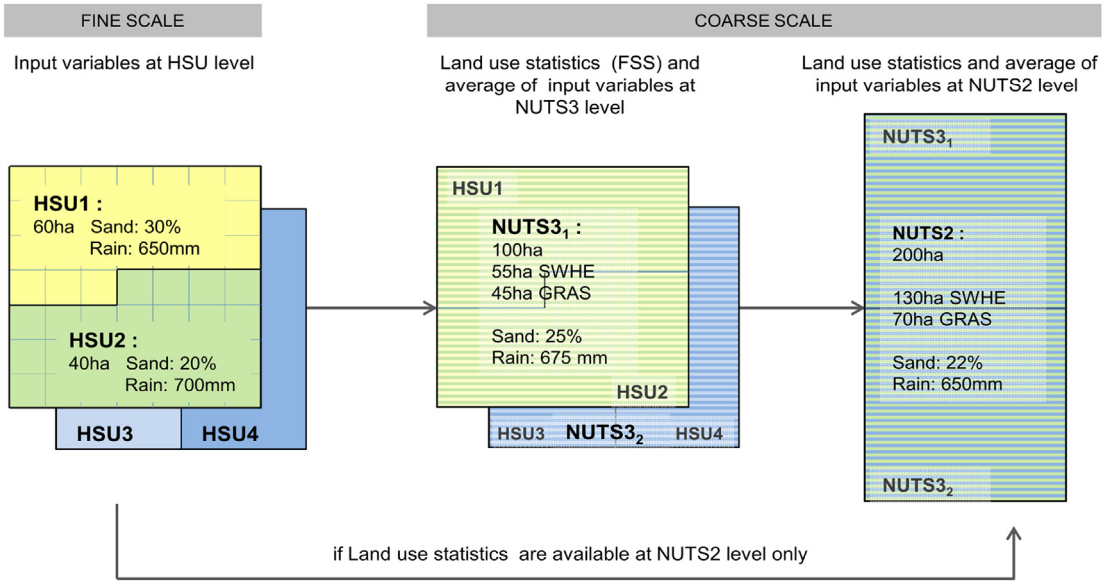

Fig. 2. Characteristics of fine-scale units and coarse-scale units and the relationship between them. A fine-scale unit is called homogeneous spatial unit and a coarse-scale unit is either NUTS3 or NUTS2 region. For instance, NUTS2 = NUTS3$_1$ + NUTS3$_2$; NUTS3$_1$ = HSU$_1$ + HSU$_2$ and NUTS3$_2$ = HSU$_3$ + HSU$_4$. Illustration of the aggregated data from fine-scale units to coarse-scale units. On one hand, input variables are available at HSU level (fine-scale). They are aggregated (mean) to get input variables at NUTS3 or NUTS2 level (coarse-scale). On the other hand, statistical data on land-use are available at NUTS3 or NUTS2 level.

and $\mathbb{I} \otimes \mathbf{x}^T$ as a matrix obtained by replacing 1 in $\mathbb{I}$ with $\mathbf{x}^T$. We use $\otimes$ as the Kronecker product.

2. Problem

In this paper we propose a methodology to predict the land-use shares or land use areas in each HSU by combining data from various sources, as illustrated in Fig. 3.

On one hand, statistical data on land-use are available at NUTS2 and NUTS3 level. On the other hand geographical information is available at high spatial resolution. The spatial and temporal variation of the geographical variables depicts the heterogeneous distribution of these variables and of land used within the administrative region. For instance, in Fig. 3, it is unlikely that the shares of grass (GRAS) and soft wheat (SWHE) are the same in both HSUs (HSU1 and HSU2) that belong to the same NUTS3 (NUTS3$_1$). We assume that part of the temporal variation in, or cross-border differences of, the land-use is also caused by price fluctuations motivating farmers to increase or reduce the cultivation of a certain crop independent of environmental conditions (Chakir, 2009).

We used the geographical, point-based observations of land-use (LU) that provide information on the repartition of land-use within an administrative region (see crosses in the box 'LUCAS data'). This information is used to link administrative and geographical information in order to characterize each point-based observation with its environmental condition (climate, soil, and land cover classes), topographical information (relief) and prices both in time and in space. With this data and the land-use statistics (FSS), the challenge

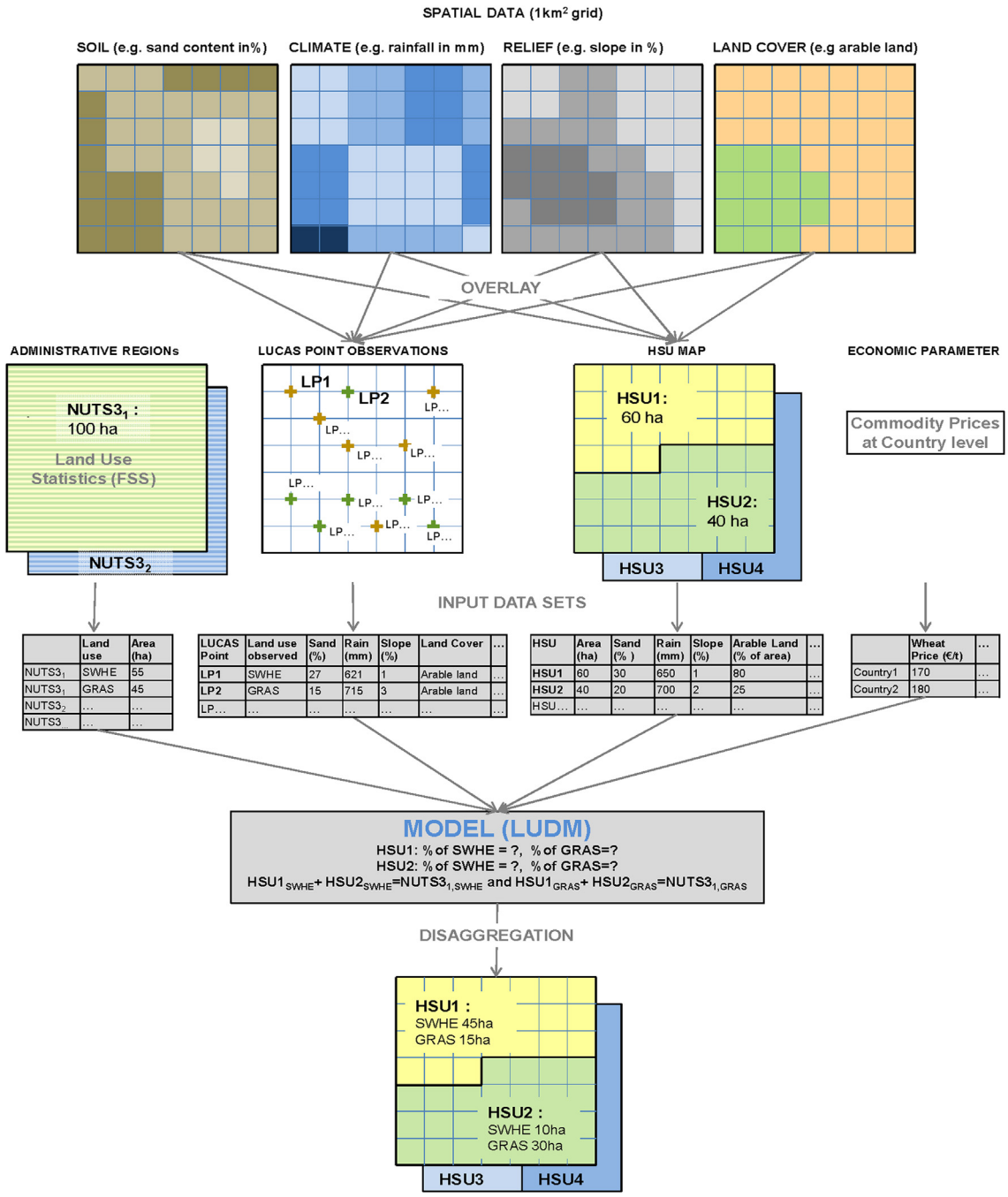

Fig. 3. Illustration of the problem for available data concerning one region. On one hand, spatial data is available at high spatial resolution. On the other hand, statistical data on land-use are available at NUTS3/NUTS2 level and price data at country level.

is to build a model (LUDM) that is able to predict the land-use areas inside each HSU and to assure that the predictions are consistent with the land-use statistics (FSS) to avoid creating or loosing the land.

3. Datasets

3.1. Land use/cover area frame survey (LUCAS) data

The LUCAS survey is a point-based field survey over EU countries and it gives, at each georeferenced location, the land-cover/use observed (for instance, see points on map d of Fig. 9). It is carried out by EUROSTAT since 2001. In this paper, we used LUCAS survey data of 2001, 2003, 2006, 2007 and 2009. The land-use classes of these data had been reclassified and regrouped into about 30 classes of agricultural land-use/cover, including forest and grassland, to match with the classification used in the CAPRI model (see Appendix J). For some of the countries, the LUCAS data was not available (e.g. Bulgaria, Cyprus, and Romania). In these cases, we used the LUCAS data from the neighboring countries.

3.2. Crop harvested and forest areas

The Farm Structure Survey (FSS, EUROSTAT, 2010) provided the observations of the agricultural land-use areas at an administrative level (NUTS3) for all the European countries except Germany for

which the data were available at only at NUTS2 level. Therefore, for Germany we had only one observation of land-use areas ($N = 1$) for each NUTS2 region, while we had at least two land use observations ($N > 1$) per NUTS2 region for the other EU-countries (see NUTS3 layer on map d of Fig. 9).

Forest areas are available at a 25 m × 25 m grid (Pekkarinen et al., 2009; Kempeneers et al., 2013) and are aggregated (summed) to the HSU level and to the regional scale (NUTS2 for Germany and NUTS3 for others EU-countries). Discussion on forest areas is available in Appendix A.

3.3. Land use/cover data

The land cover map (Coordination of information on the environment in Europe -CORINE) describes land cover based on the visual interpretation of satellite images. CORINE Land Cover (2006/2000) data are provided by the European Environment Agency (ETCSIA, 2012; ESA, 2008) for the area of EU-28. The nomenclature includes 44 classes of agricultural, urban and natural areas (see Appendix K).

3.4. Meteorological data

The meteorological data for a 25 km by 25 km grid were available from the EC-JRC AGRI4CAST (EC-JRC-AGRI4CAST, 2012). The meteorological database provided daily and interpolated data from 1975 to date for the EU-28 Member States. We used daily mean temperature and rainfall for the years 2000–2009. Discussion on data is available in Appendix A.

3.5. Biophysical data

Percentage of organic carbon content, sand and clay were based on the soil mapping units of the Harmonized World Soil Data Base-HWSD. The soil unit database content is described in detail in the documentation of the HWSD (see FAO/IIASA/ISRIC/ISS-CAS/JRC, 2009).

3.6. Topographical data

Slope and altitude for each 1 km by 1 km grid cell were based on the Digital Elevation Model (DEM) from Jarvis et al., 2008.

3.7. Price data

The prices of major agricultural products are available at a regional level (NUTS2) for different years (e.g. 2001, 2003, 2006, 2007 and 2009) from the CAPRI system. A set of selected products includes staple crops, cash crops, and animal products that can impact the choice of using land as pasture or for fodder crops.

3.8. Uncertainties in the data

To build the model, we used all these datasets from different sources and for years close to 2010 as we have the land-use statistics available for year 2010. Thus, we face with some uncertainties in data (Brodlie et al., 2012) such as:

- uncertainty in time: variation in time of some variables as we used available data for years close to 2010;
- uncertainty in the definition of variables: a variable from different datasets can have similar definition but not exactly the same;
- data confidentiality uncertainty: some values are replaced by other values for data confidentiality reasons;
- uncertainty in data source: difference among similar available datasets.

3.9. Explanatory variables

Soil characteristics (organic content, sand, and clay); topographical information (slope and altitude); meteorological variables (annual rainfall, sum of temperature and vegetation period) and CORINE land cover classes are used as explanatory variables to characterize the HSUs. In addition to these environmental variables, we add economic variables such as prices of main crops and animal products (wheat, barley, rape seeds, potatoes, milk, beef and pork) as these prices can impact the farmers' decision to grow a certain crop. We use LUCAS data from two or more NUTS2 regions or the LUCAS data for a completely different country during the process of building the model for a given NUTS2. We also use the

Table 1
Explanatory variables used in the model. We use these data for five different years corresponding to the availability of LUCAS data.

Variables (**X**)	Unit	Descriptions
Soil texture		Harmonized world soil
- Sand	%	Sand content
- Clay	%	Clay content
- OC	%	Organic carbon content
Relief		DEM data
- SLP	%	Slope
- ALT	m	Altitude
Climate		AGRI4CAST data: average of variables from 2000 to 2001; 2002–2003; 2004–2006, 2006–2007 and 2008–2009
- TEMP	C0	Annual sum of daily temperature
- RAIN	mm	Annual sum of daily rainfall
- VEGP	days	Vegetation period: total days with temperature >5 °C
Prices		Crop and meat prices from CAPRI system: prices for years 2001, 2003, 2006, 2007, 2009
- P. SUGB	Euro	Price of sugar beet
- P. COMI	Euro	Price of milk product
- P.WHEA	Euro	Price of wheat
- P.BARL	Euro	Price of barley
- P.LMAI	Euro	Price of maize
- P.RAPE	Euro	Price of rape seed
- P.POTA	Euro	Price of potatoes
- P. BEEF	Euro	Price of beef
- P. PORK	Euro	Price of pork
CORINE		Land cover classes
- CLC		14 harmonized corine classes (see Appendix K)

LUCAS data for different years. Table 1 provides an overview of the explanatory variables used.

4. Disaggregation modelling framework

This section provides the main structure of the land-use disaggregation model and the Bayesian approach used for estimating the model parameters. We start with the model structure to give an overview of the model and then present the prior and posterior estimations of the model parameters.

4.1. Global structure of the disaggregation model

The general structure of the model is outlined in Figs. 2 and 4.

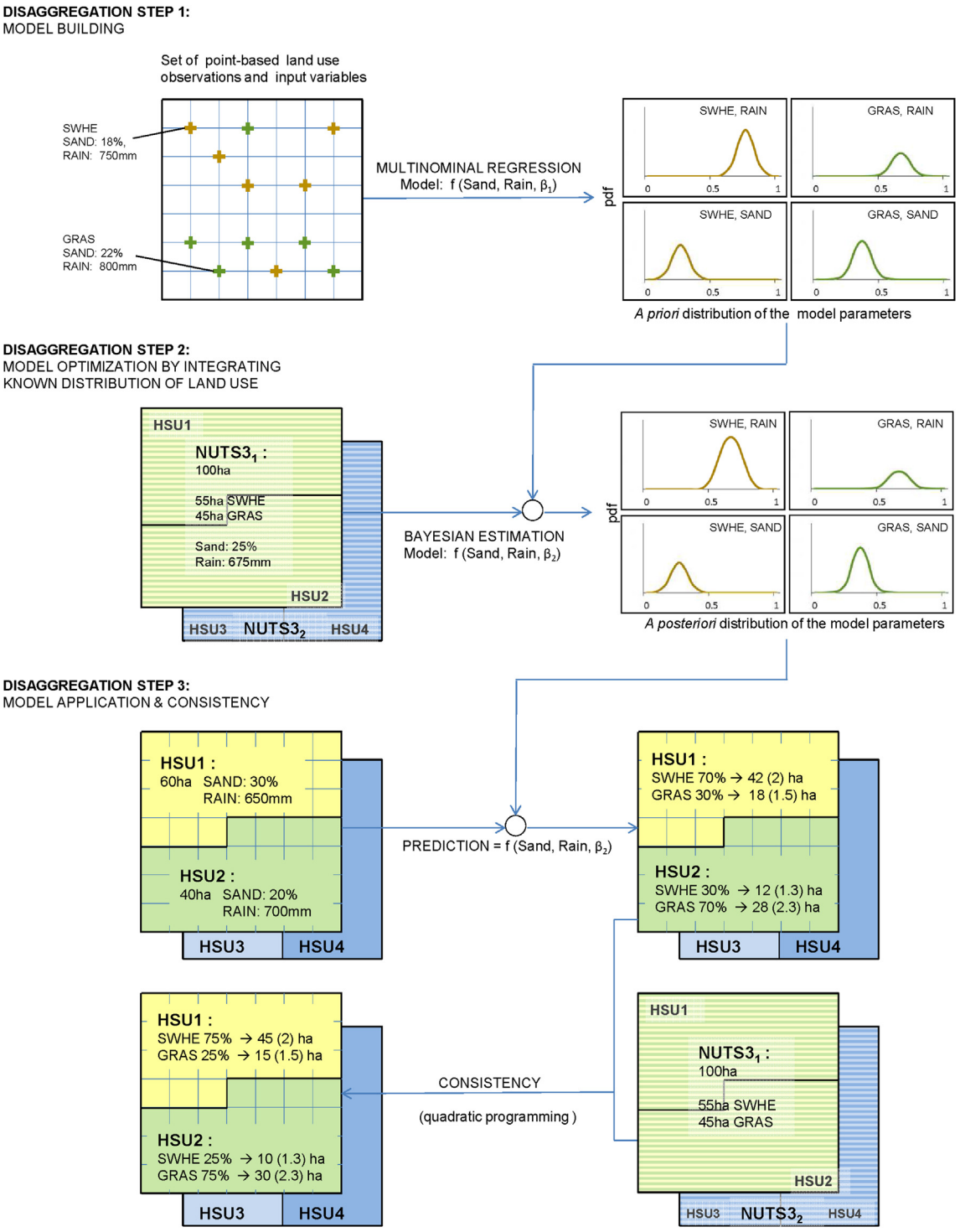

Fig. 4. Global structure of the land-use disaggregation model. Point-based observations and input variables are combined to get the *a priori* distribution of parameters (Step 1). The *a priori* distribution is combined with the land-use statistics to get *a posteriori* distribution (Step 2). The predictions of land-use areas at HSU level (using *a posteriori*) are constrained to match with both the HSU areas and the land-use statistics (Step 3).

Without loss of generality, we suppose that an administrative NUTS2 region is divided into two sub-regions ($NUTS3_1$, $NUTS3_2$) and each sub-region (for instance $NUTS3_1$) is divided into two HSUs (HSU1 and HSU2) as illustrated in Fig. 2. Fig. 2 shows that the input variables at a coarse-scale (NUTS3 or NUTS2) are the average of the input variables across all HSUs (fine-scale units) that belong to the same coarse-scale.

We assume that we have two possible land-uses (grassland-GRAS and soft wheat-SWHE); two main input variables (SAND and RAIN) that govern the allocation of agricultural land-use.

Fig. 4 shows how we use and combine the available datasets to predict the land-use areas within the HSUs. The point-based observations (LUCAS data) provide information on the repartition of land-use within the NUTS2 region. Indeed, a large land area used for a given land-use class should result in a frequent observation in the LUCAS data and vice versa. As all land-uses are in competition, given some environmental conditions (percentage of sand-SAND and annual rainfall-RAIN), we model the land-use areas (GRAS and SWHE areas) in each HSU (e.g. HSU1, HSU2) using a multinomial logit model. We assume we have the same vector of the model parameters for a given region (NUTS2). LUCAS data and the input variables are used to get the *a priori* distributions of the model parameters (β) and thus to identify the function share ($f(SAND,RAIN,\beta_1)$), which gives the proportions of land-use for a given HSU (see details in Section 5).

The function share can be used to predict the land-use areas within the HSUs belonging to the NUTS2 region using the *a priori* distributions of the model parameters (see Kempen et al., 2005; Röder and Gocht, 2013; Lamboni et al., 2013; Gocht and Röder, 2014). We consider a second step (see Section 4.3) which allows for integrating the observations of the land-use areas available mainly at a NUTS3 level into the process of estimating the model parameters.

A Bayesian approach combines the prior distributions of the model parameters and the average of input variables with the transformation of land-use areas (at NUTS3) to improve the estimations of model parameters (see details in Section 4.3). It gives the updated disaggregation function ($f(SAND,RAIN,\beta_2)$) used to get the predictions in each HSU, including the uncertainties of these predictions (standard deviation in brackets).

This step is important to:

- provide updated and improved model parameters compared to those obtained in the first step;
- correct some potential bias when working with a non-relevant point-based survey data (EU-countries not covered by LUCAS survey) or using the LUCAS data for other regions (NUTS2) or countries;
- include non-frequent land-use in the model, which are not found in the first step.

This second step furthermore allows for quickly updating the model parameters when new observations of land-use areas become available.

To be consistent with the observations of the land-uses areas, available mainly at NUTS3 level, we provide constrained predictions of the land-use areas within the HSUs using the mean square error to measure the quality of the predictions (step 3). Basically, we constrain the Bayesian predictions by minimizing the mean square error subject to some constraints (see details in Section 4.4). The constraints are: the sum of the land-use area across all HSUs must match with the area of that land-use area at a regional scale (NUTS3) and the sum of the crop area within a HSU layer must be less than the area of the HSU minus the area of the forest which is known.

4.2. The land-use disaggregation model

In each HSU, all land-uses are in a competition given the explanatory variables. A multinomial logit model (Hosmer and Lemeshow, 2000) makes use of a linear combination of all the explanatory variables (Table 1) to explain simultaneously the probabilities or the percentages of all the land-use categories using a link function: the logistic function.

Let L denote the number of all possible and exclusive land-uses within a HSU unit (h) and $\mathbf{x}_h$ be the $(d \times 1)$ vector of input variables (see Table 1); we predict the land-use area ($a_{h,l}$) for each category of land-use l using the following model:

$$a_{h,l} = \frac{\exp\left(\boldsymbol{\beta}_l^{\mathscr{T}} \mathbf{x}_h\right)}{\sum_{l=1}^{L} \exp\left(\boldsymbol{\beta}_l^{\mathscr{T}} \mathbf{x}_h\right)} \times a_h, \tag{4.1}$$

with $\boldsymbol{\beta}_l, l = 1, 2, \ldots, L$ the $(d \times 1)$ model parameters for a category of land-use l and a_h the area of the HSU unit.

Remark 4.1. *If we define* $s_{h,l} = \exp(\boldsymbol{\beta}_l^{\mathscr{T}} \mathbf{x}_h) / \sum_{l=1}^{L} \exp(\boldsymbol{\beta}_l^{\mathscr{T}} \mathbf{x}_h)$, *we have* $a_{h,l} = s_{h,l} \times a_h$. *Thus, estimating the land-use area* $a_{h,l}$ *is equivalent to estimate the share or percentage of the same land-use* $s_{h,l}$ *as the area of the HSU unit* a_h *is known.*

4.3. Bayesian estimation of the model parameters

Equation (4.1) gives the land-use area and the logarithm of the

Table 2
List of the main hypotheses and simplifications made in the processes of modelling the land-use areas inside the HSU units. As we develop one model for each administrative NUTS2-region and we cover the EU-28 countries, these hypotheses and simplifications are a compromise between the details in statistical process description and the computing capacity. They introduce some structural uncertainty in the model.

Equations	Hypothesis	Why factor
Equation 4.2	Approximation of a multivariate logistic distribution of the latent variable ($\mathbf{Z}_h$) by a multivariate normal distribution	The approximation is used to get an analytical expression of the mean and the covariance matrix of the latent variable ($\mathbf{Z}_h$) and *a posteriori* distribution in Bayesian inference (Frühwirth-Schnatter and Frühwirth, 2012)
Equation 4.3	Using a geometric mean to link the distribution of the latent variable in a given HSU layer to the distribution of the observed variable at a coarse-scale unit	In this paper, we focus on modelling the mean component of the relative proportion. A geometric mean is appropriate when working with the proportion or rate.
Equation 4.3	Independence of the latent variables that belong to the same administrative region	One of the main ideas beyond developing the HSU is that the proportions of land-use areas inside a given HSU are specific to that one and do not theoretically impact on the proportions of another HSU. Given the model parameters, $\mathbf{Z}_h$ is entirely determined by the input variables of the HSU. Therefore, we assume that the errors terms are mutually independent from one HSU to another.
Equation 4.6	Normal distribution of the empirical priors	Classic in statistical inference

relative proportion of the land-use area with respect to the referenced category (L), i.e. $\mathbf{z}_h = \left[log\left(\frac{a_{h,1}}{a_{h,L}}\right), \ldots, log\left(\frac{a_{h,L-1}}{a_{h,L}}\right)\right]^\top$, is a realization of a stochastic process $\mathbf{Z}_h$ ($\mathbf{Z}_h = \mathbf{z}_h + \boldsymbol{\epsilon}$) which follows a multivariate logistic distribution (see McFadden, 1974; Frühwirth-Schnatter and Frühwirth, 2012 for more details). A multivariate logistic distribution can be approximated by a multivariate normal distribution as follows (Kotz et al., 2005; Frühwirth-Schnatter and Frühwirth, 2012):

$$\mathbf{Z}_h \sim \mathscr{N}_{L-1}(\mathbf{X}_h\boldsymbol{\beta}, \mathbf{R}), \tag{4.2}$$

with $\boldsymbol{\beta} = [\boldsymbol{\beta}_1^\mathscr{T} - \boldsymbol{\beta}_L^\mathscr{T}, \ldots, \boldsymbol{\beta}_{L-1}^\mathscr{T} - \boldsymbol{\beta}_L^\mathscr{T}]^\mathscr{T}$ the $(d(L-1) \times 1)$ vector of parameters; $\mathbf{X}_h = \mathbb{I}_{L-1} \otimes \mathbf{x}^\mathscr{T}$ the $(L-1 \times d(L-1))$ matrix of input variables and $\mathbf{R} = (R_{i,i} = \pi^2/3, R_{i,j|i\neq j} = \pi^2/6)$ the covariance matrix of the type $(L-1 \times L-1)$ (see Kotz et al., 2005; Frühwirth-Schnatter and Frühwirth, 2012).

The likelihood of $\mathbf{Z}_h$ is used to get the Bayesian inference if the observations of the land-use areas are available at the HSU level. When the observations of the land-use areas are available at a coarse-scale (case of disaggregation), the geometric mean, convenient when working with relative proportions, allows for deriving the likelihood of the model at a coarse-scale based on the likelihood at the HSU level. Thus, the likelihood of the relative proportions of the land-use areas within a given coarse-scale unit ($\mathbf{Y}_n$) follows a multivariate normal distribution under a technical assumption of independence among the HSUs (see Table 2 and Appendix B) for more explanation and assumptions made:

$$\mathbf{Y}_n \sim \mathscr{N}_{L-1}(\mathbf{X}_n\boldsymbol{\beta}, \mathbf{R}_n), \tag{4.3}$$

with $\mathbf{X}_n = \frac{1}{H_n}\sum_{h=1}^{H_n}\mathbf{X}_h$; $\mathbf{R}_n = \mathbf{R}/H_n$ and H_n is the number of HSU units that belong to a given coarse-scale.

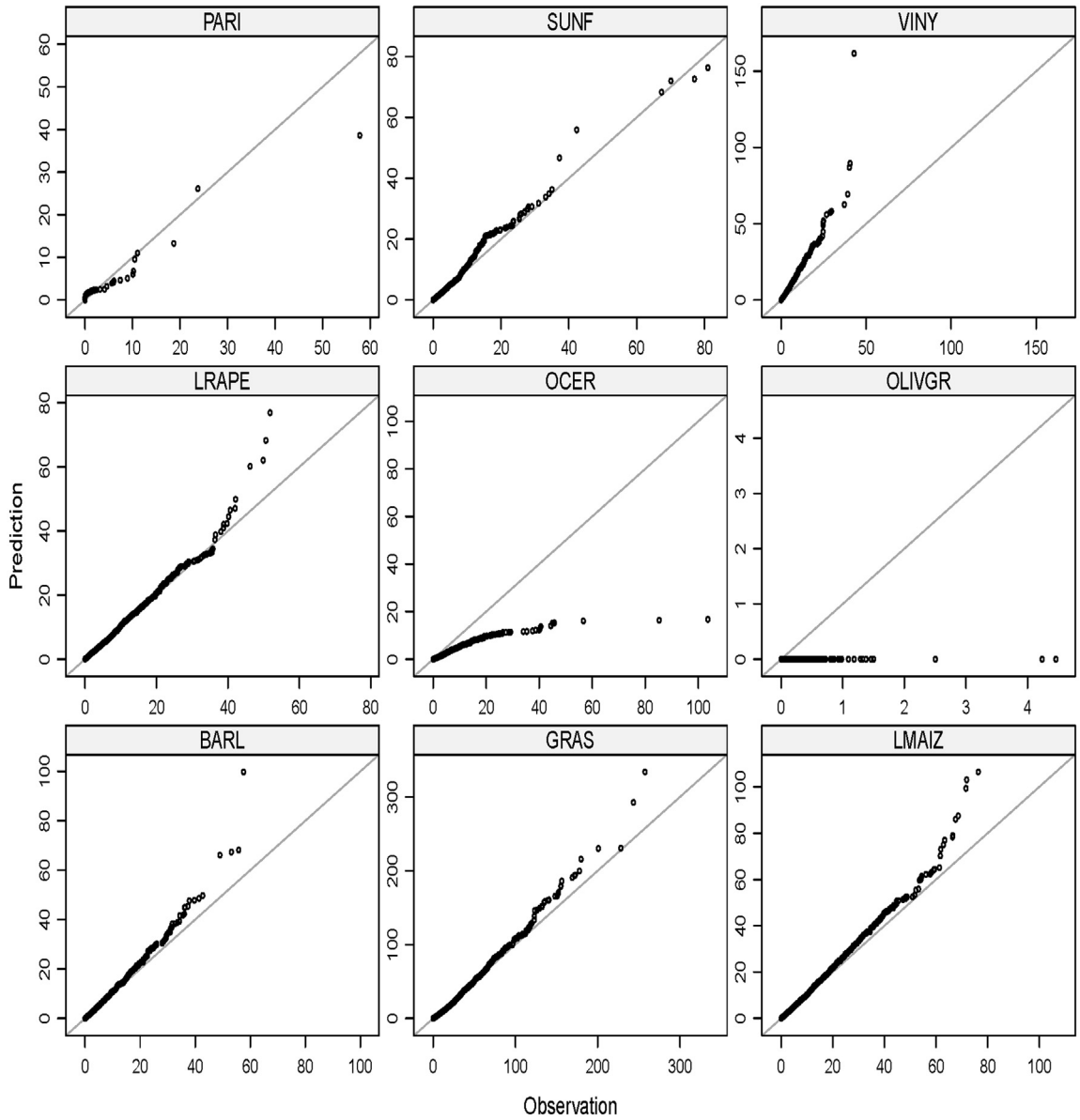

Fig. 5. Q-Q plots of the predicted land-use areas (×100 ha) and the observed land-use areas (×100 ha) at the HSUs level for France. The land-uses BARl, GRAS, LMAIZ, LRAPE, OCER, OLIVGR, PARI, SUNF and VINY stand respectively for barley, grassland, maize, rapeseed, other cereal, olive, rice, sunflowers and vineyard. The peculiar behaviours of the plots VINY, OCER, OLIVGR are mainly due to the uncertainties in the data we used for the comparison and some deviations between predicted and observed land use are evident (see paragraph 6.2).

Giving the likelihood of the model at a coarse-scale, it becomes possible to formally integrate the data of land-use areas available only at that level through a general Bayesian linear model. The general Bayesian linear model (Smith, 1973) is defined as follows:

$$\pi(\mathbf{y}_n|\boldsymbol{\beta}) \sim \mathscr{N}_{L-1}(\mathbf{X}_n\boldsymbol{\beta}, \mathbf{R}_n) \tag{4.4}$$

$$\pi(\boldsymbol{\beta}) \sim \mathscr{D}(\boldsymbol{\beta}_0, \mathbf{B}_0)\,, \tag{4.5}$$

where $\pi(\boldsymbol{\beta})$ is *a priori* distribution of the parameters with mean $\boldsymbol{\beta}_0$ and covariance matrix $\mathbf{B}_0$.

While the classical algorithm such as Metropolis-Hastings or Gibbs sampling can be used to get the posterior distributions of the parameters, we consider a conjugate *a priori*, that is:

$$\pi(\boldsymbol{\beta}) \sim \mathscr{N}_{d(L-1)}(\boldsymbol{\beta}_0, \mathbf{B}_0), \tag{4.6}$$

and the conjugate *a posteriori* distributions follow a multivariate normal distribution:

$$\pi(\boldsymbol{\beta}|\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N) = \widehat{\boldsymbol{\beta}^*} \sim \mathscr{N}_{d(L-1)}(\boldsymbol{\beta}^*, \mathbf{B}), \tag{4.7}$$

with mean and covariance matrix (Lindley and Smith, 1972; Smith, 1973; Frühwirth-Schnatter and Frühwirth, 2012):

$$\boldsymbol{\beta}^* = B\left(\mathbf{B}_0^{-1}\beta_0 + \sum_{n=1}^{N}\mathbf{X}_n^{\mathscr{T}}\mathbf{R}_n^{-1}\mathbf{y}_n\right) \tag{4.8}$$

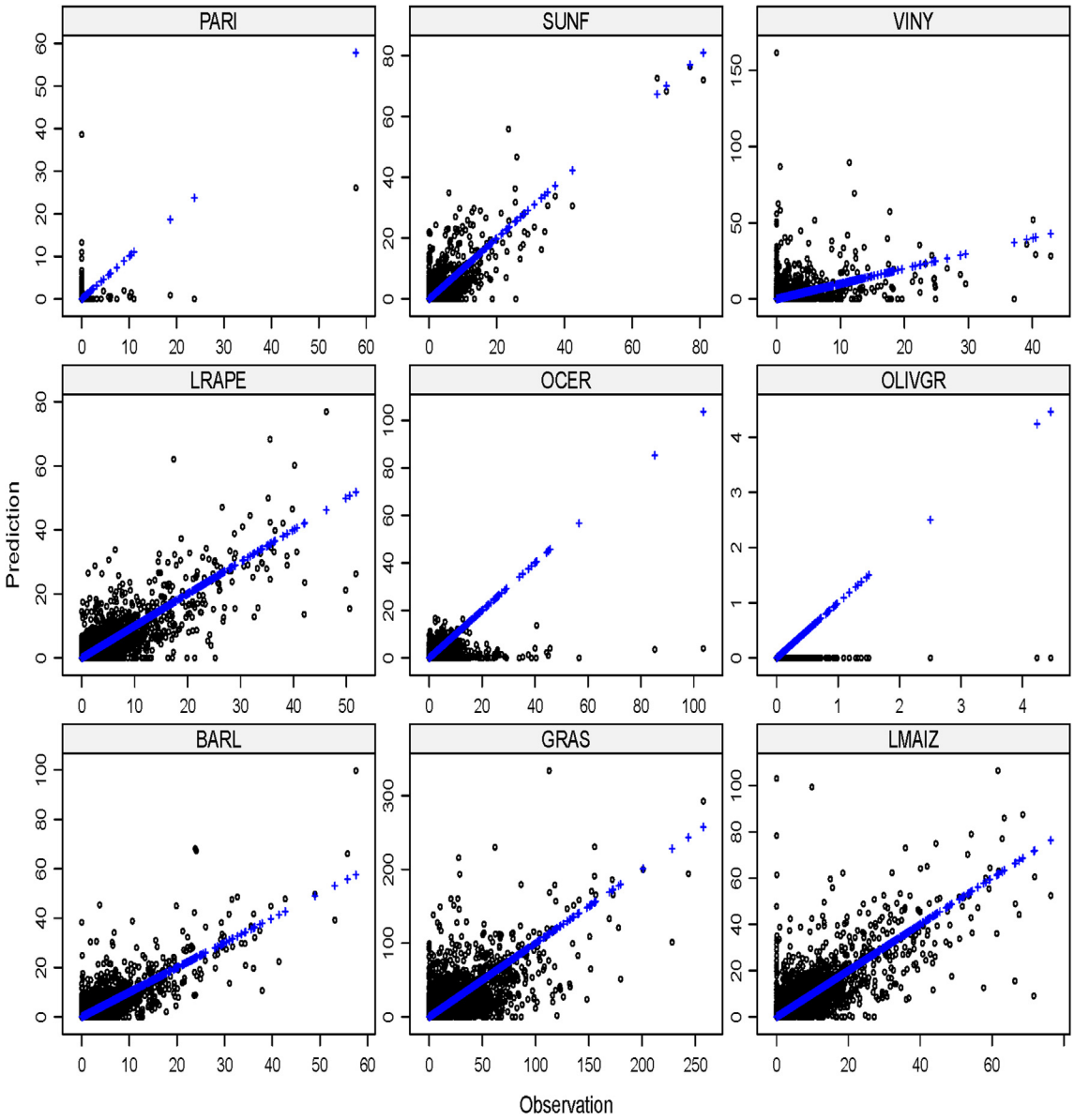

Fig. 6. Scatter plots of the predicted land-use areas (×100 ha) and the observed land-use areas (×100 ha) at the HSUs level for France. The black points are the scatter plots of the prediction versus the observation and the blue ones are the plot of the observation versus the observation. The land-uses BARl, GRAS, LMAIZ, LRAPE, OCER, OLIVGR, PARI, SUNF and VINY stand respectively for barley, grassland, maize, rapeseed, other cereal, olive, rice, sunflowers and vineyard. The peculiar behaviours of the plots PARI, VINY, OCER, OLIVGR are mainly due to the uncertainties in the data we used for the comparison and some deviations between predicted and observed land use are evident (see paragraph 6.2). (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this article.)

$$\mathbf{B} = \left(\mathbf{B}_0^{-1} + \sum_{n=1}^{N} \mathbf{X}_n^T \mathbf{R}_n^{-1} \mathbf{X}_n \right)^{-1}. \quad (4.9)$$

where N is the number of the observations of the land-use areas available at a coarse-scale. In our case, it represents the number of land-use areas available in a given administrative region (NUTS2 region). The $(L-1 \times 1)$ vector $\mathbf{y}_n, n = 1,2,\ldots,N$ is the logarithm of the proportion of each observation of the land-use area with respect to the area of the referenced category. The matrix $\mathbf{X}_n$ is the input variable related to the observation $\mathbf{y}_n$. It is obtained as an average of the input variables for all the H_n-HSUs that belong to the sub-region where the observation $\mathbf{y}_n$ has been made (see equation (4.3)).

We summarize and motivate the assumptions and the approximations made in this section in Table 2. With regard to the invariance of natural patterns in the environment, the hypotheses in Table 2 are equivalent to assume that the logarithm of the relative proportions are scale invariant with respect to all the input variables used in this paper. Indeed, given the model parameters, the transformation of the land-use areas in a given HSU $\mathbf{z}_h$ is linear with respect to all the inputs $\mathbf{X}_h$ and it is known that a linear function is scale invariant (see Mandelbrot, 1983; Sellers et al., 1997; Li, 2000). Moreover, the scale invariant properties come from the distribution properties of $\mathbf{Z}_h$ in equation (4.2) and $\mathbf{Y}_n$ in equation (4.3) which is the mean of $\mathbf{Z}_h$ across a given region and both have the same distribution (see Li, 2000).

4.4. Optimal and constrained predictions of the land-use areas

Based on equations (4.1) and (4.7), the unconstrained Bayes predictor of the land-use share $\widehat{S_{h,l}}$ (see remark 4.1) is:

$$\widehat{S_{h,l}} = \frac{\exp\left(\widehat{\boldsymbol{\beta}^*}_l^{\mathcal{T}} \mathbf{x}_h\right)}{\sum_{l=1}^{L} \exp\left(\widehat{\boldsymbol{\beta}^*}_l^{\mathcal{T}} \mathbf{x}_h\right)}. \quad (4.10)$$

Let $\widehat{\mathbf{S}_\mathbf{l}} = [\widehat{S_{1,l}}, \ldots, \widehat{S_{h,l}}, \ldots, \widehat{S_{H_n,l}}]^T$ be a $(H_n \times 1)$ vector of the Bayes predictors of the land-use shares across all the H_n HSU units that

Fig. 7. Summaries (mean, third quartile and median) of the error terms (in ×100 ha) of the predictions at HSU level using equation (6.15) for the first 11 NUTS2 regions in France (FR10, FR21, …, FR43). The land-uses BARl, GRAS, LMAIZ, LRAPE, OCER, OLIVGR, PARI, SUNF and VINY stand respectively for barley, grassland, maize, rapeseed, other cereal, olive, rice, sunflowers and vineyard.

belong to a given coarse-scale unit. Let $\widehat{\mathbf{S}} = [\widehat{\mathbf{S_1}}^T, \ldots, \widehat{\mathbf{S_l}}^T, \ldots, \widehat{\mathbf{S_{L-1}}}^T]^T$ be a $((L-1)H_n \times 1)$ vector of all the Bayes predictors of land-use shares. The vector $\widehat{\mathbf{S}}(\widehat{\beta^*})$ is a random vector and the optimal and unconstrained predictor of the shares is $\mathbb{E}[\widehat{\mathbf{S}}|\widehat{\beta^*}]$ when we consider the general mean square error to measure the performance of a predictor. The general mean square error of $\widehat{\mathbf{S}}$ is defined as $\mathbb{MSE}(\widehat{\mathbf{S}}) = \mathbb{E}[(\widehat{\mathbf{S}} - \mathbf{s})^T \mathbf{W}(\widehat{\mathbf{S}} - \mathbf{s})]$, with $\mathbf{W}$ a symmetric matrix and $\mathbf{s}$ the true and unknown shares.

In practice, we would require minimize the mean square error under some constraints on the predicted shares. In the case of land-use, the most important constraints are:

- **C1**: the weighted sum of the shares of a land-use (l) across all HSUs must match with the area of that land-use observed at a coarse-scale unit ($A_{l,n}$) with $l = 1,2,\ldots,L-1$ and $n = 1,2,\ldots,N$;
- **C2**: the sum of the crop shares within a HSU must be less than $1 - s_{h,fore}$, with $s_{h,fore}$ the known percentage of the forest as we can also have other land-use type L;
- **C3**: the shares must be positive to avoid getting some negative land-use area.

Let $\mathbf{S}$ be a vector of constrained shares of $\widehat{\mathbf{S}}$. We can summarize all possible and compatible constraints in one equation like:

$$\mathbf{CS} - \mathbf{c} \geq 0 \tag{4.11}$$

where $\mathbf{C}$ is the constraint matrix (see Appendix C), containing the weights applied for constraining the predictors and the row number of $\mathbf{C}$ is the number of constraints. We use $\mathbf{c}$ as the constrained values, mainly the shares available at a coarse-scale. For more detail see Appendix C.

The following proposition provides the optimal and constrained predictor of the shares.

Proposition 4.1. *Assume that $\widehat{\mathbf{S}}$, $\mathbf{S}$ have a second moment; $\mathbf{S}$ like $\widehat{\mathbf{S}}$ is a function of $\widehat{\beta^*}$; $\mathbf{CS} - \mathbf{c} \geq 0$ is a set of compatible constraints and the constraint region is bounded. If we use the generalized mean square error as the performance criterion of a predictor, we have:*

(i) *the conditional, optimal and constrained predictor of the shares ($\widetilde{\mathbf{S}}$) is the solution of the quadratic programming (QP) problem:*

Fig. 8. Summaries (mean, third quartile and median) of the error terms (in $\times 100$ ha) of the predictions at HSU level using equation (6.15) for the last 11 NUTS2 regions in France (FR51, FR52, …, FR83). The land-uses BARI, GRAS, LMAIZ, LRAPE, OCER, OLIVGR, PARI, SUNF and VINY stand respectively for barley, grassland, maize, rapeseed, other cereal, olive, rice, sunflowers and vineyard.

$$\tilde{\mathbf{S}} = \underset{\mathbf{S}}{\operatorname{argmin}} \frac{1}{2}\mathbf{S}^{\mathscr{T}}\mathbf{W}\mathbf{S} - \mathbf{S}^{\mathscr{T}}\mathbf{W}\hat{\mathbf{S}} \tag{4.12}$$

$$\textit{subject to} \quad \mathbf{C}\mathbf{S} - \mathbf{c} \geq 0. \tag{4.13}$$

(ii) *If* $\mathbf{W}$ *is a positive definite matrix, we have a strictly convex QP problem and this problem has an unique and global solution.*

Proof 4.1. *See proof in* Appendix D.

5. Implementation issue

5.1. Computation of the empirical priors

In this paper, we use a local multinomial logit regression to get the empirical priors of the model parameters and, as in classical analysis, we assign a multivariate normal distribution for the

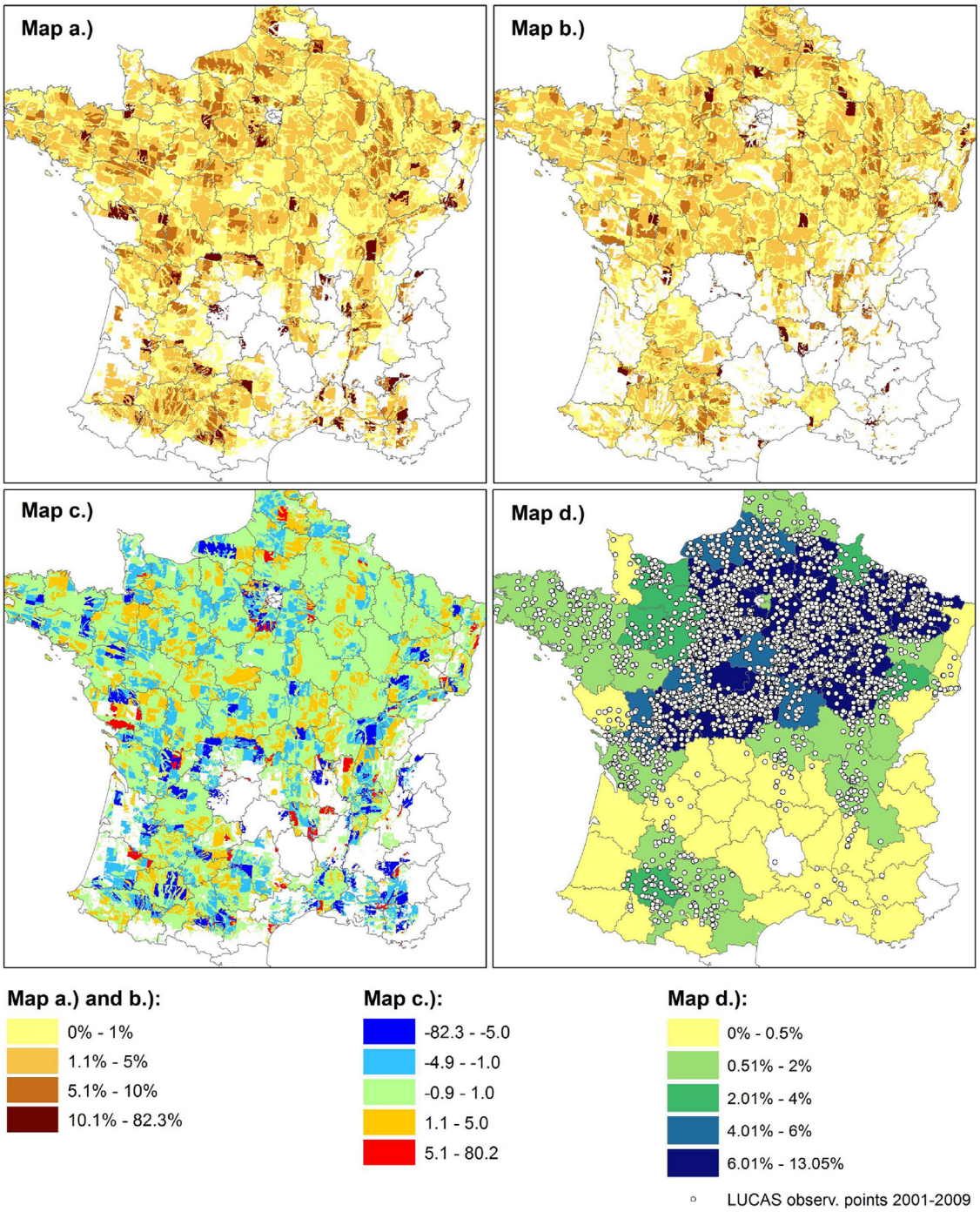

Fig. 9. Comparison of French LPIS data and the constrained disaggregation results for rapeseed for the year 2010. The upper maps show rapeseed area in the HSU as % of total rapeseed area in the NUTS3 region based on French LPIS data (Map a.) and from the disaggregation (Map b.). Map c. gives the difference between LPIS data and the disaggregation results (Difference between Map b. and Map a.). Map d. presents the original FSS data at NUTS3 level (rapeseed as % of total NUTS3 area) and the LUCAS rapeseed observations (points) in the years 2001–2009.

estimator of the model parameters. The local multinomial logit regression allows for preserving the (local) environmental properties and it consists in regressing the vector of point-based observations on the explanatory variables by giving high weight to the observations close to the center of a desirable administrative region (NUTS2 in this paper) and low weight to other points (for more details see Appendix E, Lamboni et al., 2013, 2014a; Lamboni, 2013; Lamboni et al., 2014a). The weight is based on i) the distance between the center of the NUTS2 and the point-based observation; ii) a potential bandwidth, i.e. the maximum distance for which the weight is not zero.

The set of the potential bandwidths (distance values), used to build the model at each NUTS2 region, are the same for each country and are listed in Table E.3 (Appendix E) according to our previous tests. The optimal bandwidth for a given NUTS2 region is chosen among the set of potential bandwidths. The optimal bandwidths for two or more NUTS2 regions can be different. Table E.3 in Appendix E reports the minimum and the maximum values of the optimal bandwidths determined by the model for each country. These optimal bandwidths are used to identify the LUCAS point-based observations and the associated input variables to be included in the model and then to get the prior estimates of the model parameters. For a selected optimal bandwidth (optimal LUCAS data), we assign a non-informative prior estimates for all the possible land-use classes which are not found in that LUCAS data. For these land-use classes, we assigned zero to the means of *a priori* distributions of the model parameters and we used the highest value of the standard deviations found in this region as standard deviations.

All the LUCAS data and the associated input variables from 2001 up to 2009 were used to get *a priori* distributions of the model parameters.

5.2. Computation of the Bayesian estimates

During the second step of the Bayesian estimation, we used the input variables for 2009 and the land-use areas for 2010 (FSS data). Based on formulas in (4.8), each observation of land-use areas at a NUTS3 level $\mathbf{A}_n = ([A_{1,n}, \ldots, A_{l,n}, \ldots, A_{L,n}]^T)$ is transformed to get: $\mathbf{y}_n = \left[\log\left(\frac{A_{1,n}}{A_{L,n}}\right), \ldots, \log\left(\frac{A_{l,n}}{A_{L,n}}\right), \ldots, \log\left(\frac{A_{L-1,n}}{A_{L,n}}\right)\right]$. The associated input variables at NUTS3 level ($\mathbf{X}_n$) are the average of the input variables at the HSU level ($\mathbf{X}_h$).

For Germany, we did the same but at NUTS2 level as the observations of land-use areas are available at a NUTS2 level.

Notice that we did not use the direct land-use areas ($\mathbf{A}_n$) but a transformation of these areas. We used 10^{-4} as zero ha to avoid the logarithm transformation of values that include zero. The Bayesian estimates of the model parameters were obtained by applying the

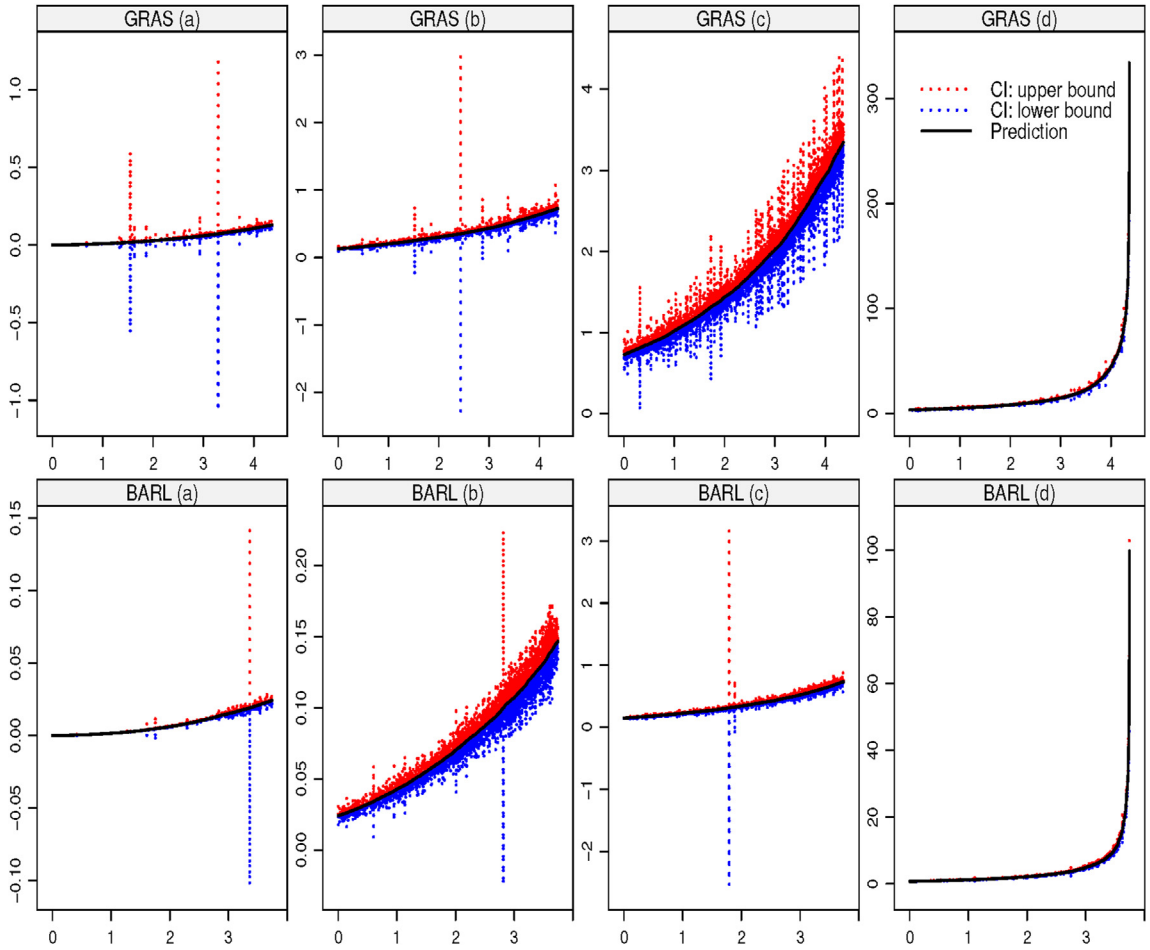

Fig. 10. Uncertainties (95 % confident intervals in ×100 ha) of predictions at HSU level after sorting and dividing all the HSUs into four groups according to the predicted areas of a given land-use and their quartiles. We distinguish four panels (e.g. BARL (a), BARL (b), BARL (c), BARL (d)) corresponding to the four groups. The land-uses BARL, GRAS stand respectively for barley, grassland (see paragraph 6.3). The *x*-coordinate reports the number of HSUs divided by 1000.

formulas in (4.8).

5.3. Computation of the predictions

Given the Bayesian estimates of the model parameters and the input variables at a HSU level, we ran the model (equation (4.1)) to get the predictions of the land-use shares or areas within the HSUs.

5.4. Computation of the constrained predictions

In this step, we used the direct land-use areas ($\mathbf{A}_n$) and the predictions from the model (equation (4.1)). We constrained the predictions in each HSU according to the Proposition 4.1 using the observed $\mathbf{A}_n$ of the NUTS3 region for all EU countries except Germany. For Germany, we used $\mathbf{A}_n$ of the NUTS2 region.

5.5. Computation of the predictions uncertainties

To get the predictions uncertainties, we have to run the model for different values of the model parameters choosing within the posterior distributions of model parameters (equation (4.7)).

To better sample within the parameters space, we firstly selected the most influential model parameters using sensitivity analysis (Saltelli et al., 2008; Lamboni et al., 2008, 2009, 2011b) as we have about 900 parameters in a NUTS2 region. The total sensitivity index of a given parameter ($\beta_{j,l}$, $j = 1,\ldots,d$ and $l = 1,\ldots,L-1$) for the linear model with multiple outputs ($\mathbf{Y}_n = [Y_{n,1} = \mathbf{x}_n^{\mathcal{T}}\boldsymbol{\beta}_l, \ldots, Y_{n,L-1} = \mathbf{x}_n^{\mathcal{T}}\boldsymbol{\beta}_{L-1}]^{\mathcal{T}}$) in equation (4.3) is (Lamboni et al., 2008, 2009, 2011a):

$$S_{T_{\cdot j,l}} = \frac{\sum_{l=1}^{L-1} x_{n,j}^2 \mathbb{V}\mathrm{ar}\left(\beta_{j,l}\right)}{\sum_{l=1}^{L-1} \mathbb{V}\mathrm{ar}(Y_{n,l})}, \tag{5.14}$$

where $x_{n,j}$ is the input variable associated to parameter $\beta_{j,l}$ for a category of land-use l; $\mathbb{V}\mathrm{ar}(\beta_{j,l})$ is the variance of $\beta_{j,l}$ and $\mathbb{V}\mathrm{ar}(Y_{n,l})$ is the variance of the output $Y_{n,l}$.

Secondly, we sampled 50 values of he most influential parameters (about 5%) within the posterior distributions of these parameters and we fixed the other parameters to their mean values in (4.7) (Lamboni et al., 2009, 2014b).

Thirdly, we ran the model for each of the sample values and computed the confident interval of the predicted land-use areas within HSUs (Lamboni et al., 2014b).

6. Disaggregation results and discussion

In this section, we discuss the predicted results of the land-use areas in order to give some evidences about the level of accuracy of

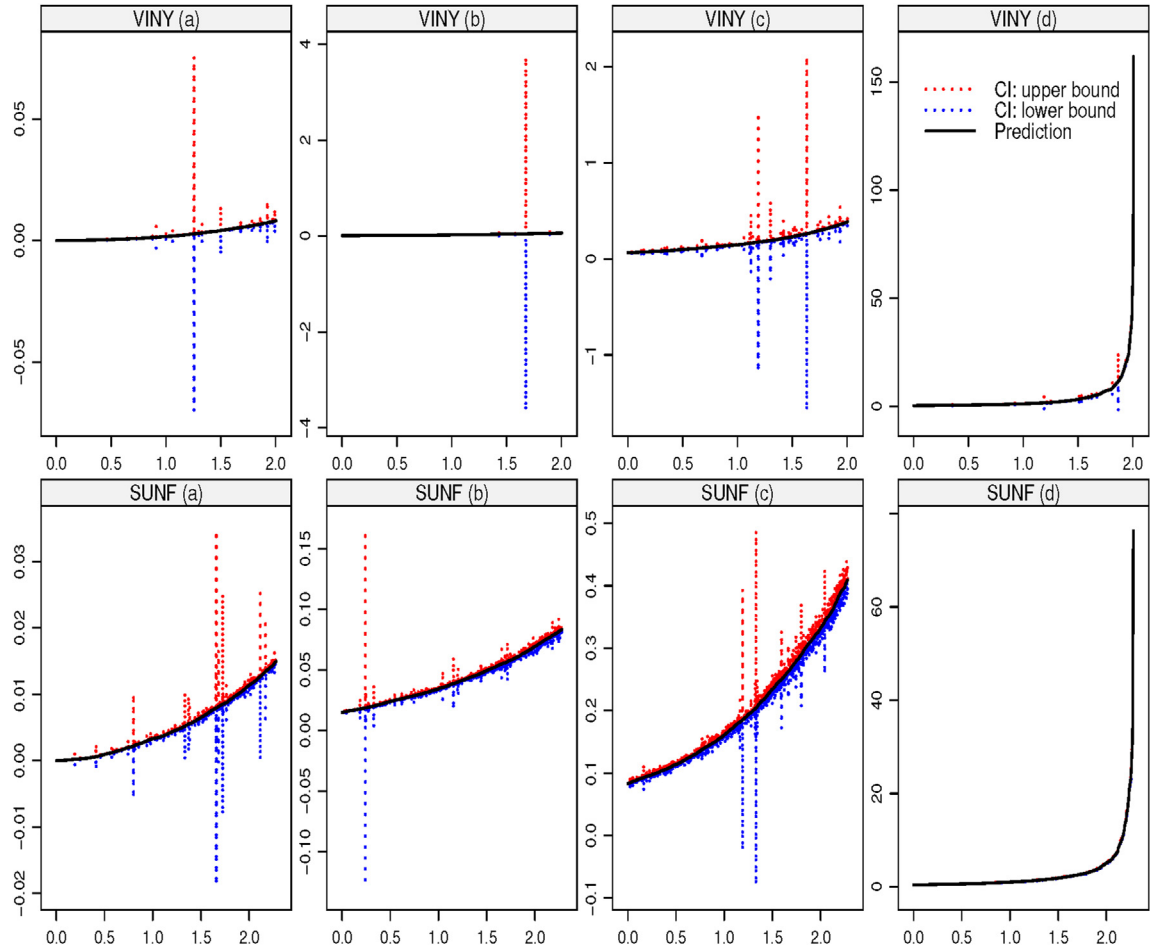

Fig. 11. Uncertainties (95% confident intervals in $\times 100$ ha) of predictions at HSU level after sorting and dividing all the HSUs into four groups according to the predicted areas of a given land-use and their quartiles. We distinguish four panels (e.g. SUNF (a), SUNF (b), SUNF (c), SUNF (d)). The land-uses SUNF and VINY stand respectively for sunflowers and vineyard (see paragraph 6.3). The *x*-coordinate reports the number of HSUs divided by 1000.

the model predictions and to draw some perspectives for further improvement of the model. First, we show that our predictions match with the FSS data as we are going to use these data for the validations and for some comparisons; and second, we assess the quality of the results for France for which data on crop areas at the HSU level are available (Cantelaube and Carles, 2015), and third, we represent the spatial distribution of barley to illustrate the results. We also compare our results to those obtained from few years ago (Kempen et al., 2005; Leip et al., 2008) to discuss some benefits and disadvantages of both results.

As expected, the predicted land-use areas are consistent with the FSS data. Appendix F shows the quantiles of the predicted areas versus observed areas (Q-Q plots) at NUTS2 (resp. NUTS3) level for 30 possible land-uses in Germany (resp. France).

6.1. Results

To check the quality of the predictions, we received independent data from the Land Parcel Identification System (LPIS) for France (Cantelaube and Carles, 2015). The LPIS data were re-mapped from parcel level into the HSU. Fig. 5 shows the quantiles of the predicted areas versus the LPIS areas at the HSU level for the 9 land-uses available in the LPIS dataset. While the predictions for more frequent crops such as barley, maize, grassland, rapeseed and sunflower match with the French LPIS data, we observe some miss-predictions for non-frequent crops like vineyard (VINY), rice (PARI) and olive. Going into more detail, Fig. 6 shows the scatter plots of the predicted versus LPIS land-use areas at the HSU level. It confirms the previous results, i.e. the model predicted better the areas of the main crops than the less frequent. Moreover, for the main crops, the model miss-predicted the highest values of the areas compared to the others values probably due to the presence of the outliers (see the maximum values of the errors in Fig. 8).

We use the weighted predictor error (from Chakir, 2009) to account for the agricultural area of the NUTS3-regions (with the observations of land-use areas) as a small error does not have the same meaning in a NUTS3-region with big area and in a NUTS3-region with small area. The errors terms are calculated as follows:

$$E_{l,h} = \sqrt{\left(\widehat{a_{h,l}} - a_{h,l}\right)^2 \times \frac{a_h}{\sum_{h=1}^{H_n} a_h}}, \tag{6.15}$$

with $\widehat{a_{h,l}}$ (resp. $a_{h,l}$) the predicted (resp. observed) land-use area; a_h the area of a HSU and H_n the number of HSUs in a given region (NUTS3).

Figs. 7 and 8 provide the summary of these error terms for respectively the first 11 NUTS2 and the last 11 NUTS2 regions in France. We focused on the median of the errors in our analysis to avoid the effect of the outliers which affect both the mean and the maximum values. More details about the summaries are in Appendix G.

Bearing in mind that the smallest HSU has an area of 100 ha, the predictor errors less than 1.9 ha for barley across the 22 regions (highest median error found in region FR10) show that in more than 50% of the HSUs, the model predictions for barley can be considered good. For several regions (FR42, 61, 71, 81, 82, 83) more than 75% (third quartile) of the errors for barley are less than 1.9 ha.

For grassland, the maximum of the median values of the errors across the 22 regions is about 12.4 ha (obtained in FR63) and the minimum value of the median is 0.53 ha (from FR10). The model has better performance in predicting grassland than barley in the region FR10 and vice versa in the 21 regions.

The areas of maize are also well predicted across the regions with the maximum of the median errors of 3.59 ha (in FR42) and the minimum of zero (FR81-83). In general, the errors of other cereals (OCER) are small: maximum median-value of 1.5 ha and most of the third quartile values less than 4 ha. In the case of rice (PARI) and olive (OLIVGR), the maximum of the third-quartile values is less than 0.2 ha (obtained in FR83). But we have some outliers in FR82 with 117.3 ha for olive and 1069.4 ha for rice. The model faces also some difficulties for vineyard (VINY) in FR82 where the maximum value reaches 2007.07 ha (see Appendix G).

Fig. 9 shows the spatial distribution of the share of rapeseed for both the observations (map a) and the predicted results (map b) at the HSU level. Map c is the error map (difference between map a and map b). In order to better understand our results, map d gives the locations where rapeseed areas are important (Center-North of France). Overall, the disaggregation results are close to the observations in most of the locations. In the Center-North of France, the crop distribution is predicted quite well for the majority of NUTS3 regions as the errors are lower in these regions.

However, it is challenging to allocate a small crop area at the exact location of the observation within a NUTS3 region as it is the case for Central-Southern of France. We can see some gaps that occur mainly in some NUTS3 regions with small areas of rapeseed (30 ha). For these NUTS3 regions, the disaggregation model assigns the total area of rapeseed (in the NUTS3 region) to only a few small HSU.

6.2. Effect of the uncertainties in the data on the prediction

While these comparisons are necessary to give a level of confidence in using the model predictions, we have to keep in mind also the uncertainties in the data we used for the comparison. For data confidentiality reasons, FSS reports data only if a certain minimum number of farms are included in the data sets. Therefore, non-frequent crops might be reported with zero area even though the crop is cultivated in the region. For example, FSS data reports zero area for olive in some NUTS3 regions. In this case, we predicted zero area of olive in all the HSU units due to the constraints we applied. For rice (PARI), FSS data reported cultivations in only three NUTS3 regions. Moreover, the aggregated (summed) areas of other cereals (OCER) from the LPIS data do not match with the areas reported in the FSS data at the NUTS3 level, probably due to the uncertainties in i) the definition of other cereals; ii) splitting out some LPIS parcel into two or more HSU units when a LPIS parcel intersects these HSUs.

6.3. Predictions uncertainties

Of course, the predictions uncertainties of a non grown crop in a NUTS3/2 region (zero area) is null. As a matter of fact, we leave out these crops in our analysis (Figs. 10 and 11). Figs. 10 and 11 show 95% confident intervals of the predictions at HSU level for respectively barley, grassland and sunflowers, vineyard (see Appendix H for other land-uses).

The confident bounds are generally proportional to the level of the predictions and are close to the predictions. High uncertainties have mainly found in panels land-use (a, b) and low uncertainties in land-use (c, d). We found that in a very small number of HSUs (Fig. 10 BARL (c) and GRAS (b), Fig. 11 VINY (b, c) for instance), the predictions are not precise.

However, the predictions uncertainties depend on i) the number of the model runs; ii) the threshold used to select the number of model parameters (see sensitivity indices in Appendix I) and the independent assumption in sensitivity analysis.

6.4. Discussion

The process of the prediction of land-use areas involves a combination of different input variables such as altitude, slope, rain and land cover (CORINE) classes. While the rain can be more important for growing cereals for instance, it can be less important in the case of rice (as the rice field is often irrigated) and in the case of some permanent crops like olive. Our model predicts some small area of cereals like maize or wheat also at altitudes which are not suitable for growing cereals in Europe. Equally, the model predicts small area of rice when the slope is relative high. These results seem to be unrealistic as the rice field requires flat terrain for irrigation purpose. Although we have distinguished the individual CORINE classes for rice and olive as explanatory variables, the model still faces some difficulties to better predict these non-frequent crops. These miss-predictions are likely due to: i) using only one set of model parameters at NUTS2 level to predict all the results in different locations (HSUs) of the NUTS2 region. Indeed, rice, olive and wheat can be seen in one sub-region (NUTS3 for instance) and not at all in others sub-regions of the same NUTS2 region; ii) not including, in this paper, agronomic constraints for the suitability or limits of growing certain crops under certain environmental conditions; iii) using a land cover classes (corine) of year 2006; iv) using a non-informative prior for the crops which are not found in the first step.

6.5. LUDM versus a rule-based approach

Under the assumptions of no sub-NUTS3 heterogeneity, we use a uniform distribution as a rule-based approach (Thomas-Agnan and Vanhemsz, 2013) to distribute the land-use areas across the HSUs. Fig. 12 shows the errors distributions for both the LUDM and the simple rule-based (RB) approaches. It comes from Fig. 12 that the LUDM results outperform the RB results for land-uses barley,

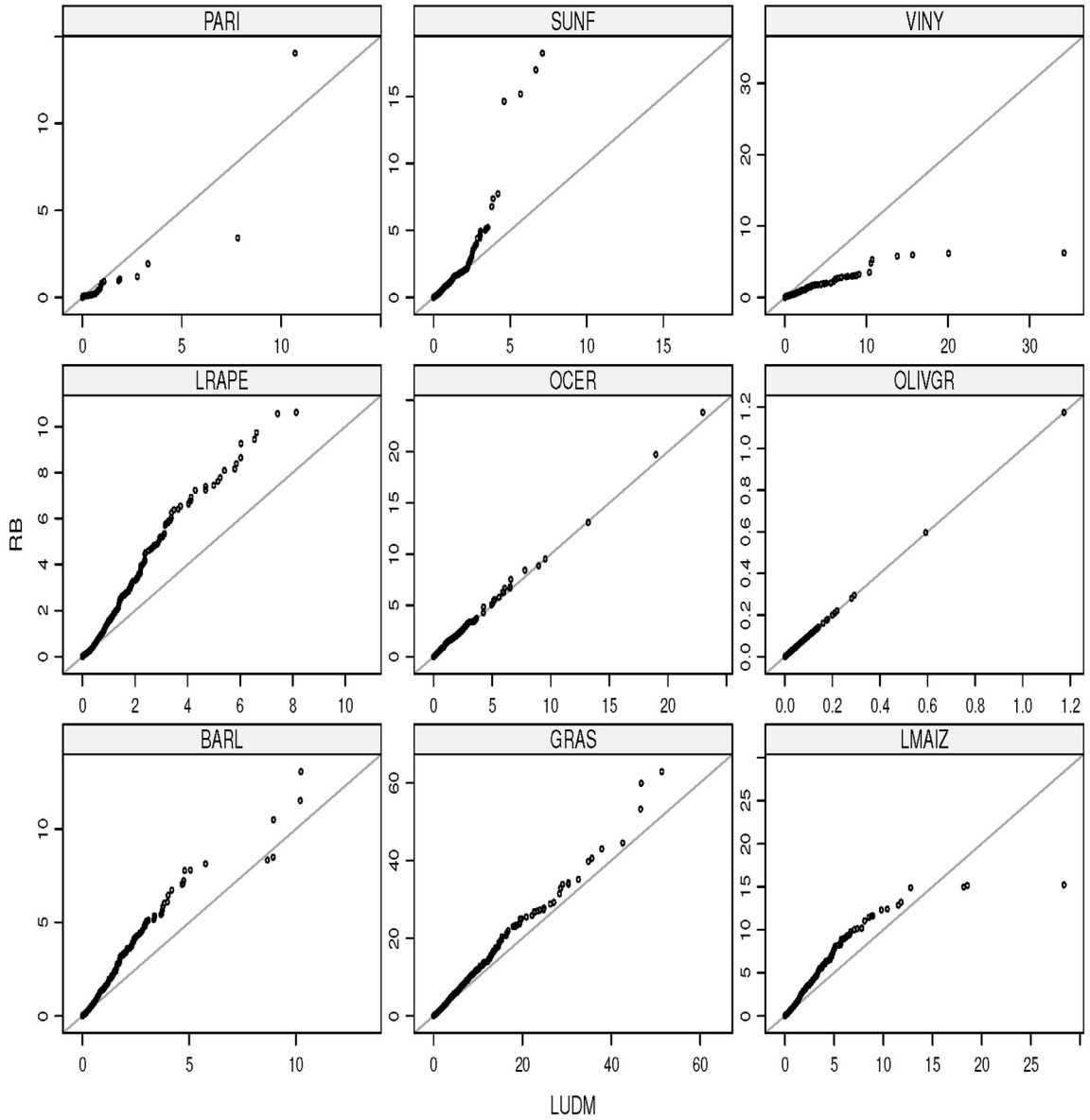

Fig. 12. Q-Q plots of the errors terms in (6.15) (×100 ha) for both the LUDM and the rule-based (RB) approaches at the HSUs level for France. The land-uses BARl, GRAS, LMAIZ, LRAPE, OCER, OLIVGR, PARI, SUNF and VINY stand respectively for barley, grassland, maize, rapeseed, other cereal, olive, rice, sunflowers and vineyard (see paragraph 6.5).

grassland, rapeseed, sunflower and maize. For other cereals, we have approximately the same distributions of errors across HSUs and the RB approach performs better in the case of rice and vineyard. The case of olive is obvious (see Section 6.2).

6.6. LUDM results of barley for 2010 versus results of barley for 2000

Fig. 13 shows a comparison between the predicted spatial distributions of barley shares in 2010 over EU-28 countries and the results obtained for the year 2000 for EU-15 (Leip et al., 2008). The main differences between the two maps are: first, different input data used (data round the year 2000 used by (Leip et al., 2008) and data around the year 2010 in this study; second the definition of the spatial units differed. Leip et al., 2008 included the CORINE classes in the delineation of the spatial units while in our approach CORINE is used as explanatory variable in the model. Furthermore, in Leip et al., 2008 no meteo-grid was used in the delineation and no 'no-go' areas were identified allowing also larger spatial units. Third, Leip et al., 2008 used the dasymetric approach predicting crop by crop with independent binomial logit models in the first step for EU-15 countries while this study uses a multinomial logit model.

Nevertheless, barley distribution is similar (only in location) over EU-15 countries.

7. Conclusion

In this paper, we were interested in predicting land-use areas inside the homogenous spatial units (HSUs) over EU-28 countries

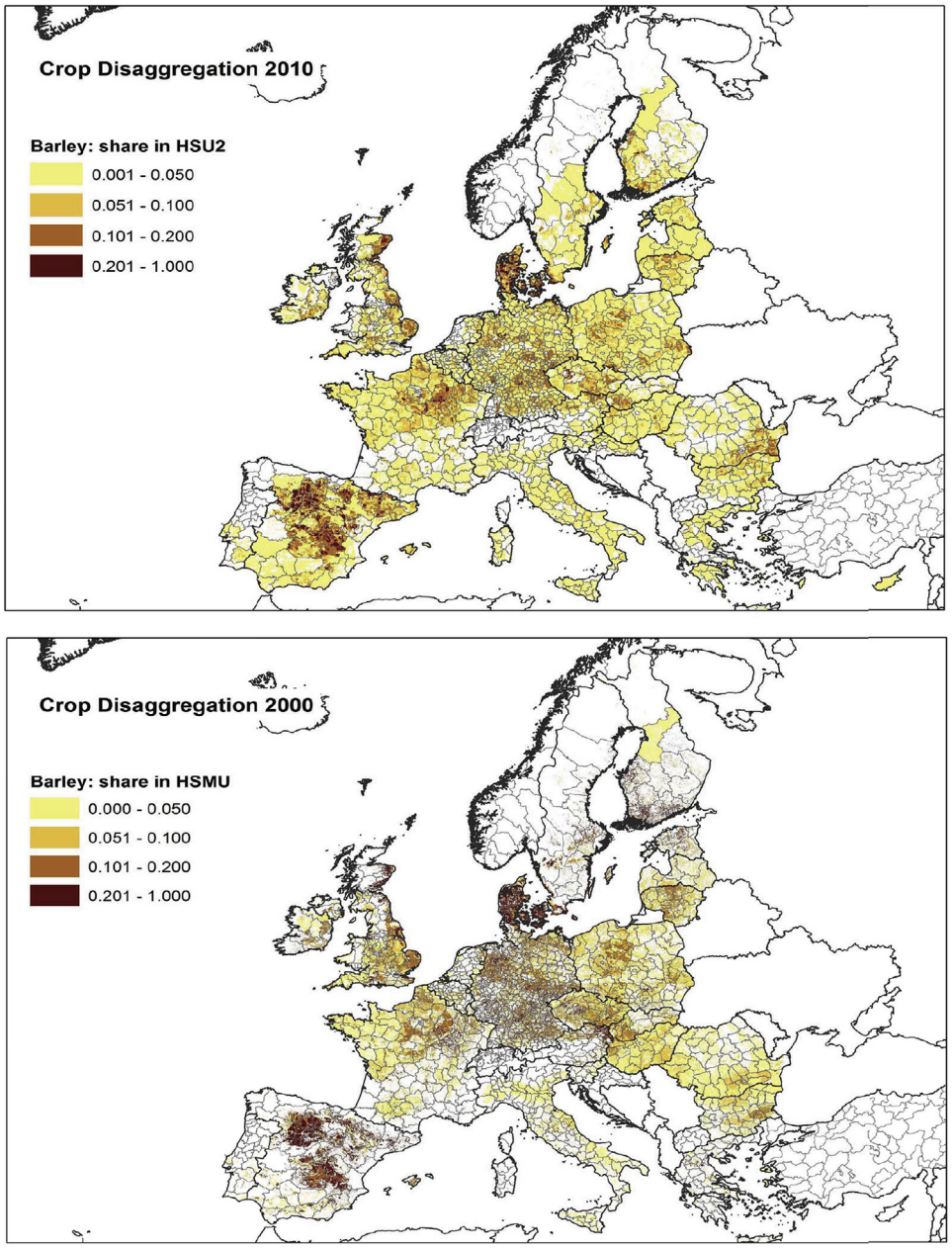

Fig. 13. Spatial distributions of barley shares in 2010 and in 2000. The map for the year 2000 is adapted from Leip et al., 2008.

by combining data from the Land Use/Cover Statistical Area Frame Survey (LUCAS) data and the Farm Structure Survey (FSS) data. The predictions of the land-use areas inside the HSUs are based upon multinomial models built for each administrative NUTS2 region. An optimal Bayesian approach is used to get the estimates of the model parameters and the constrained predictions in order to avoid creating or losing the land area. Due to the scale invariant properties of the model, it can be adapted for different scales based on the availability of the data. For example, if an administrative data at a finer level than NUTS3 becomes available. A known land-use class (such as forest areas) (will) serve as a referenced category to get the relative proportions in the process of deriving the land-use areas at a scale of interest.

The predicted results are consistent with the FSS data available at NUTS2/3 level and the comparison with LPIS data for France gives confidence in the accuracy of our predictions for the main land-uses while caution is needed when using the predicted areas of non-frequent crops. The current results are affected by i) the uncertainty in input data and in model structural uncertainty (O'Hagan, 2012; Beven and Freer, 2001; Rinderknecht et al., 2012; Lamboni et al., 2014b); ii) numerical approximations (using 10^{-4} ha as zero ha in logarithm transformation of land-use areas); iii) software uncertainty (de Rigo, 2013: Lehman and Belady, 1985; Lehman, 2000; Dönmez and Grote, 2011). The predictions serve as an approximation for deriving reliable indicators of environmental impact assessment. Moreover, the model can be used for future predictions and associated indicators of land-use for the HSUs.

In this paper, the model predictions of the land-use areas are generally accurate and outperform the results of a simple rule-based approach for frequent land-uses. But the model miss-predicts the highest values due to the presence of the outliers. The model is developed by focusing on the mean component of the land-use areas and there is a need to handle these outliers. One way to deal with this issue is to increase the number of observation of the land-use areas to better cover some large NUTS2-regions. While we proposed one model (parameters) for each administration region NUTS2 according to the data available, it might be possible to significantly increase the level of accuracy of the predictions by developing e.g. one model per each sub-region NUTS3 and by including agronomic constraints regarding the explanatory variables. If the methodology were to be applied for NUTS3 regions, more detailed data (data available at a level lower than NUTS3) were required. In Europe, such detailed land-use information is existing in many countries, but not available for research purposes due to data confidentiality issues. If the land-use disaggregation model could be built with the help of such detailed data, derived agri-environmental indicators could gain in accuracy in both ex-post and ex-ante assessments of agricultural policies.

Acknowledgments

We thank the four reviewers for their careful reading and detailed comments that have helped improving this paper. Special thanks to one of the reviewers for helping to improve the readability of the paper by readers experts in different domains. Many thanks to MARS team for providing the data and others.

Appendix A. General discussion on data

Appendix A.1. Choice of land use/cover and forest data sets

At the beginning of this study (i.e. 2012), the CORINE land/use cover 2006 was the most recent high resolution land use/cover data set available for Europe. To ensure temporal consistency we chose a high resolution forest map covering the same period (reference year 2006). To date CORINE land/use cover 2012 map as final product with partial validation (http://land.copernicus.eu/pan-european/corine-land-cover/clc-2012) and the Forest Type 2012 map, as partially validated product (http://land.copernicus.eu/pan-european/high-resolution-layers/forests/forest-type/view), are available.

Appendix A.2. Choice of meteorological data set

The EC-JRC AGRI4CAST gridded meteo data set is operationally used in the Crop Yield Forecasting System (MCYFS) at the European Commissions Joint Research Centres Monitoring Agricultural Resources Unit (EC JRC MARS). MCYFS provides information on crop production of the current growing season for the European Commissions implementation of the Common Agricultural Policy (CAP). The gridded meteo data set is continuously developed further regarding interpolation of underlying meteorological sites, spatial resolution etc. Our final choice to use the EC-JRC AGRI4CAST meteo data set was driven by the fact that besides temperature and precipitation also radiation, evapotranspiration, wind speed and snow depth are available. For the disaggregation itself only temperature and precipitation is taken into account. Though, for future applications based on crop information disaggregated to the HSU level also other parameters (e.g. evapotranspiration to calculate the water balance) will be of interest.

However there are other meteo data sets freely available to the research community. For example the European Climate Assessment & Dataset project provides daily gridded meteo information at a similar resolution. The E-OBS data set (Haylock et al., 2008) is currently limited to temperature, precipitation and sea level pressure information. The data set is available at http://eca.knmi.nl/download/ensembles/download.php.

Both, EC-JRC AGRI4CAST gridded meteo data and E-OBS data have their weaknesses and strengths. One of the weaknesses of the EC-JRC AGRI4CAST gridded meteo data is the interpolation scheme of precipitation while long-term temperature is represented quite well in time and space (Andrea Toreti, pers. comm. Jan. 2015). A discussion on uncertainties of the Ensembles E-OBS data set is available in Hofstra et al., 2009.

Appendix B. Likelihood of the Model

It comes from Equation (4.2) that

$$\mathbf{Z}_h \sim \mathcal{N}(\mathbf{X}_h\boldsymbol{\beta}, \mathbf{R}), \tag{B.1}$$

where $\mathbf{R} = (R_{i,i} = \pi^2/3, R_{i,j|i\neq j} = \pi^2/6)$ and $\mathbf{X}_h = \mathbb{I} \otimes \mathbf{x}^T$ is a matrix of input variables.

The quantity $(\exp(\mathbf{Z}_h))$ follows a multivariate log-normal distribution. The geometric mean of $\exp(\mathbf{Z}_h), h = 1, 2, \ldots, H_n$ over all the H_n-HSUs (h) within a given NUTS2 follows a multivariate log-normal distribution if we assume the independence between $\mathbf{Z}_h$, $h = 1,2\ldots,H_n$. By taking the logarithmic of the geometric mean of $\exp(\mathbf{Z}_h), h = 1, 2, \ldots, H_n$ within a given NUTS2 $(\mathbf{Y}_n)$ we have:

$$\mathbf{Y}_n = \frac{1}{H_n}\sum_{h=1}^{H_n}\mathbf{Z}_h, \tag{B.2}$$

and it follows a multivariate normal distribution:

$$\mathbf{Y}_n \sim \mathcal{N}(\mathbf{X}_n\boldsymbol{\beta}, \mathbf{R}_n), \tag{B.3}$$

with $\mathbf{X}_n = \frac{1}{H_n}\sum_{h=1}^{H_n}\mathbf{X}_h$; and $\mathbf{R}_n = \mathbf{R}/H_n$.

Appendix C. Constraints

Let $\mathbf{S}_l = [S_{1,l}, \ldots, S_{h,l}, \ldots, S_{H_n,l}]^T$ denote the $(H_n \times 1)$ vector of the shares of land-use with $l = 1,2,\ldots,L-1$ across the H_n HSUs; $\mathbf{a} = [a_1, \ldots, a_h \ldots, a_{H_n}]^{\mathscr{T}}$ be the $(H_n \times 1)$ vector of the HSU areas and $A_{l,n}$, $l = 1,\ldots,L-1$ be the $L-1$ observation of the land-use areas at a given coarse-scale. Let $0^{\mathscr{T}} = [0, 0, \ldots, 0]$ be a vector of size H_n; $1\mathrm{I}_i^{\mathscr{T}} = [0, \ldots, 0, 1, 0, \ldots 0]$ be a vector of size H_n with all components equal to 0 except the i^{th} component which is 1. We use $\mathbf{s}_{fore}$ as the vector of the known shares of forest for all the H_n HSUs.

The practical constraints listed in Section 4.4 and divided into three blocks, corresponding respectively to the constraints **C1**, **C2**, **C3** are formally defined in the matrix $\mathbf{C}$ of type $(L-1+H_n+(L-1)H_n) \times ((L-1)H_n)$; $\mathbf{W} = \mathbb{I}_{(L-1)H_n}$; $\mathbf{c}$ a vector of size $(L-1+H_n+(L-1)H_n)$ and we have:

$$\mathbf{C} = \left(\begin{array}{cccc} \mathbf{a}^{\mathscr{T}} & \mathbf{0}^{\mathscr{T}} & \ldots & \mathbf{0}^{\mathscr{T}} \\ \mathbf{0}^{\mathscr{T}} & \mathbf{a}^{\mathscr{T}} & \ldots & \mathbf{0}^{\mathscr{T}} \\ \ldots & \ldots & \ddots & \ldots \\ \mathbf{0}^{\mathscr{T}} & \mathbf{0}^{\mathscr{T}} & \ldots & \mathbf{a}^{\mathscr{T}} \\ \hline -1\mathrm{I}_1^{\mathscr{T}} & -1\mathrm{I}_1^{\mathscr{T}} & \ldots & -1\mathrm{I}_1^{\mathscr{T}} \\ -1\mathrm{I}_2^{\mathscr{T}} & -1\mathrm{I}_2^{\mathscr{T}} & \ldots & -1\mathrm{I}_2^{\mathscr{T}} \\ \vdots & \vdots & \vdots & \vdots \\ -1\mathrm{I}_{H_n}^{\mathscr{T}} & -1\mathrm{I}_{H_n}^{\mathscr{T}} & \ldots & -1\mathrm{I}_{H_n}^{\mathscr{T}} \\ \hline 1\mathrm{I}_1^{\mathscr{T}} & \mathbf{0}^{\mathscr{T}} & \ldots & \mathbf{0}^{\mathscr{T}} \\ \vdots & \vdots & \vdots & \vdots \\ 1\mathrm{I}_{H_n}^{\mathscr{T}} & \mathbf{0}^{\mathscr{T}} & \ldots & \mathbf{0}^{\mathscr{T}} \\ \mathbf{0}^{\mathscr{T}} & 1\mathrm{I}_1^{\mathscr{T}} & \ldots & \mathbf{0}^{\mathscr{T}} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0}^{\mathscr{T}} & 1\mathrm{I}_{H_n}^{\mathscr{T}} & \ldots & \mathbf{0}^{\mathscr{T}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}^{\mathscr{T}} & \mathbf{0}^{\mathscr{T}} & \ldots & 1\mathrm{I}_1^{\mathscr{T}} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0}^{\mathscr{T}} & \mathbf{0}^{\mathscr{T}} & \ldots & 1\mathrm{I}_{H_n}^{\mathscr{T}} \end{array}\right) \quad \mathbf{S} = \left(\begin{array}{c} \mathbf{S}_1 \\ \vdots \\ \mathbf{S}_l \\ \vdots \\ \mathbf{S}_{L-1} \end{array}\right) \quad \mathbf{c} = \left(\begin{array}{c} A_{1,n} \\ \vdots \\ A_{l,n} \\ \vdots \\ \hline \frac{A_{L-1,n}}{-1+s_{1,f}} \\ \vdots \\ -1+s_{h,f} \\ \vdots \\ -1+s_{H_n,f} \\ \hline \mathbf{0}_1 \\ \vdots \\ \mathbf{0}_l \\ \vdots \\ \mathbf{0}_{L-1} \end{array}\right)$$

with $\mathbf{0}_1 = \mathbf{0}_l = \mathbf{0}_{L-1} = \mathbf{0}$, $n = 1,2,\ldots,N$. We have the equalities in the first block and the inequalities in the remaining blocks.

Appendix D. Proof of the Proposition

Using the Pythagoras theorem, we have:

$$\mathbb{MSE}\left(\mathbf{S}\middle|\widehat{\beta^*}\right) = \mathbb{MSE}\left(\widehat{\mathbf{S}}\middle|\widehat{\beta^*}\right) + \mathbb{E}\left[\left(\mathbf{S}-\widehat{\mathbf{S}}\right)^{\mathscr{T}}\mathbf{W}\left(\mathbf{S}-\widehat{\mathbf{S}}\right)\middle|\widehat{\beta^*}\right] \tag{D.1}$$

As we cannot reduce $\mathbb{MSE}(\widehat{\mathbf{S}}|\widehat{\beta^*})$, we have to minimize the last term in (D.1) subject to $\mathbf{C}S - \mathbf{c} \geq 0$. Given a value of $\widehat{\beta^*}$ (estimation of β) The conditional expectation:

$$\mathbb{E}\left[\left(\mathbf{S}-\widehat{\mathbf{S}}\right)^{\mathscr{T}}\mathbf{W}\left(\mathbf{S}-\widehat{\mathbf{S}}\right)\middle|\widehat{\beta^*}\right] = \left(\mathbf{S}-\widehat{\mathbf{S}}\right)^{\mathscr{T}}\mathbf{W}\left(\mathbf{S}-\widehat{\mathbf{S}}\right) \tag{D.2}$$

$$\mathbb{E}\left[\left(\mathbf{S}-\widehat{\mathbf{S}}\right)^{\mathscr{T}}\mathbf{W}\left(\mathbf{S}-\widehat{\mathbf{S}}\right)\middle|\widehat{\beta^*}\right] = \mathbf{S}^{\mathscr{T}}\mathbf{W}\mathbf{S} - 2\mathbf{S}^{\mathscr{T}}\mathbf{W}\widehat{\mathbf{S}} + \widehat{\mathbf{S}}^{\mathscr{T}}\mathbf{W}\widehat{\mathbf{S}} \tag{D.3}$$

The proof of the point (ii) can be found in Goldfarb and Idnani, 1983.

Appendix E. Empirical priors: range of bandwidths and of optimal bandwidths

The local multinomial logit regression consists in regressing the vector of point-based observations on the explanatory variables ($\mathbf{x}$) by giving high weight to the observations close to the center of a desirable administrative region (NUTS2 in this application) and low weight to other points. Multinomial logit regression aims at estimating the probability to find a land-use (l) at point i ($\mathbb{P}_r(p_i = l \mid \mathbf{x})$) and is defined as follows:

$$\mathbb{P}_r(p_i = l|\mathbf{x}) = \frac{\exp\left(\beta_l^{\mathscr{T}}\mathbf{x}_i\right)}{\sum_{l=1}^{L}\exp\left(\beta_l^{\mathscr{T}}\mathbf{x}_i\right)}, \tag{E.1}$$

where, β_l is the $(d \times 1)$ vector of the parameters for the land-use l and $\mathbf{x}_i$ is the $(d \times 1)$ vector of input variables observed at point i. The vectors of model parameters β_l, $l = 1,2,\ldots L$ are estimated by maximizing the weighted log-likelihood ($\log\mathscr{L}$):

$$\log\mathscr{L} = \sum_{i=1}^{M}\sum_{l=1}^{L} w_i(c)\log[\mathbb{P}_r(p_i = l|\mathbf{x})]1\mathrm{I}_{\{p_i=l\}}, \tag{E.2}$$

with w_i, the weight given to observation i; M, the total number of observations (point-based observations) and $1\mathrm{I}_{\{p_i=l\}} = 1$ if resource-use l is found at point i and 0 otherwise.

We use the tricube weight function

$$w_i(c) = \left[1 - \left(\frac{d(c,i)}{d_c}\right)^3\right]^3 1\mathrm{I}_{\{d(c,i)<d_c\}}, \tag{E.3}$$

to calculate the weight of each observation as it decreases smoothly to zero (Cleveland, 1979; Cleveland and Devlin, 1988) when the observation at point i is so far from the center of a region (c) with respect to the bandwidth (d_c). We use $d(c,i)$ as the distance between the center c of the region and point i, and $1\mathrm{I}_{\{d(c,i)<d_c\}}$ is equal to 1 if $d(c,i)<d_c$ and 0 otherwise. However, we gave 1 as the weight for the points inside the disk of radius d_h and inside the same region (NUTS2) in order to account for some administrative constraints and other economic conditions. For a selected bandwidth, the multinomial regression provides the mean and the covariance matrix of the model parameters. Classically, we assign a multivariate normal distribution for the estimator of the model parameters. For the land-uses which are not found inside the disk of radius the selected bandwidth, we fixed the mean at zero and the standard deviation at the highest value of the standard deviations found in the multinomial regression. As we built as many models as the number of region (NUTS2) in each country, these choices were motivated by the fact that we need to be close to the behavior of the model in the same region.

The bandwidth is important as it should determine whether the observations are sufficient to represent the real feature or not. When we have to include the explanatory variables used to delineate the HSUs in the regression model, we need at least a bandwidth that ensures that almost all these variables are informative,

i.e. their variances are not close to zero. Moreover, in case of multinomial logit regression, we want our model to predict and to reproduce as well as possible the observation of land-uses/covers. These considerations motivate the use of the cross-validation approach (Stone, 1974; Yang, 2007) to choose the bandwidth. The criterion of the model quality, used in the cross-validation, is based upon the F-measure or F-score (Powers, 2011). The F-score measures how well the multinomial model is able to reproduce and to predict any land-use/cover. The F-measure is a harmonic mean between the precision and recall or sensitivity. The precision (P), for land-use l, is the rate of the well predicted (WP) l among the total prediction (TP) of l, that is $P = WP/TP$. The recall (R) of sensitivity is the rate of the well predicted (WP) l among the total observations (TO) of l: $R = WP/TO$. So, the F-measure or F-score is between 0 and 1 and is defined as follows:

$$F = 2 \times \frac{P \times R}{P + R}. \tag{E.4}$$

For all possible land-use classes, the accuracy reported in this paper is the average of the individual F-scores, i.e. the micro-average.

Based on our previous tests, the range of the potential bandwidths are given in Table E.3 and the optimal bandwidths are obtained using the cross-validation approach. We split the dataset into five groups $G_j, j = 1\ldots5$ with the same number of observations at least for the first four groups. For group G_1, first, data $G_j, j = 2\ldots5$ served to estimate the multinomial logistic model, and secondly, group G_1 is used to predict the land-use/cover and to calculate the F-measure F_1. We replicated this process for groups $G_j, j = 2\ldots5$, and the final F measure is an average of the five F-measures $F_j, I = 1\ldots5$ in order to account for the variability within the data. The bandwidth with the maximum F-measure is the optimal bandwidth among the potential ones. The optimal bandwidth is used to calculate the final weights and to estimate the parameters of the model.

Table E.3
Bandwidths for EU-28 countries. The potential bandwidths are a set of values regularly chosen between the minimum (Min) and the maximum (Max) with the step (Step). The optimal bandwidths are selected by the model among the potential ones. As we can have different optimal bandwidths for different regions in the same country, we provide the minimum (Min) and the maximum (Max) of the optimal values. All the values are in km.

Potential bandwidths			Optimal bandwidths		
Countries	Min	Max	Step	Min	Max
Austria	20	150	5	20	40
Belgium	10	150	5	20	60
Bulgaria	300	480	10	300	340
Croatia	50	150	5	135	135
Cyprus	600	700	10	660	660
Czech Republic	20	150	5	20	45
Denmark	50	150	5	50	50
Estonia	180	350	10	280	280
Finland	50	150	5	50	85
France	50	150	5	50	75
Germany	25	150	5	25	60
Greece	50	150	5	50	130
Hungary	50	150	5	50	50
Ireland	50	250	10	60	200
Italy	50	150	5	50	65
Latvia	50	150	5	75	75
Lithuania	50	150	5	60	60
Luxembourg	50	150	5	50	50
Malta	25	150	5	100	100
Netherlands	35	150	5	35	70
Poland	50	150	5	50	50
Portugal	50	150	5	50	85
Romania	200	450	10	200	450
Slovakia	50	150	5	50	60
Slovenia	50	150	5	55	55
Spain	50	150	5	50	80
Sweden	50	150	5	50	80
United Kingdom	15	150	5	15	60

Appendix F. Comparison between predicted land-use areas and FSS data

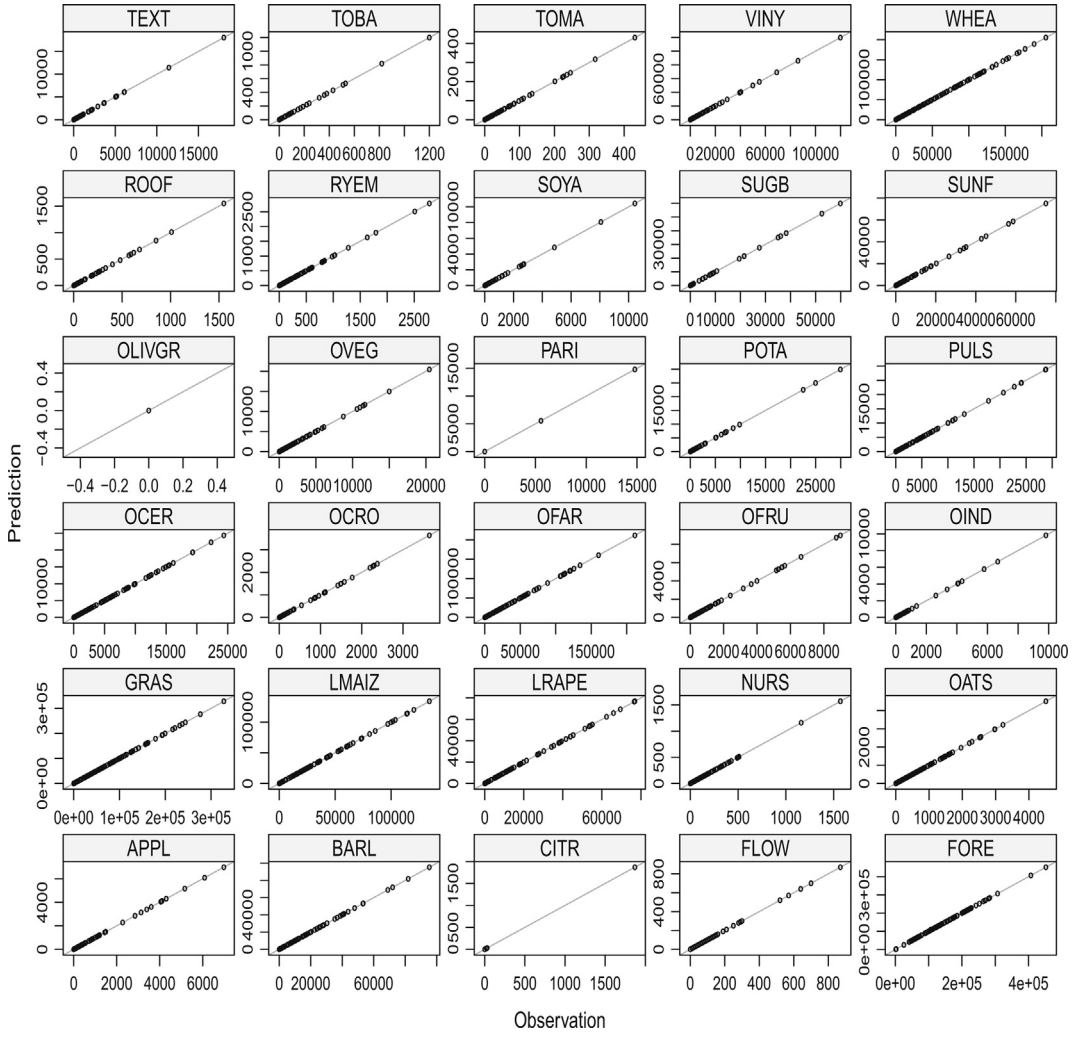

Fig. F.14. Q-Q plots of aggregated predictions of land-use areas (in ha) versus FSS observations at NUTS3 level for France.

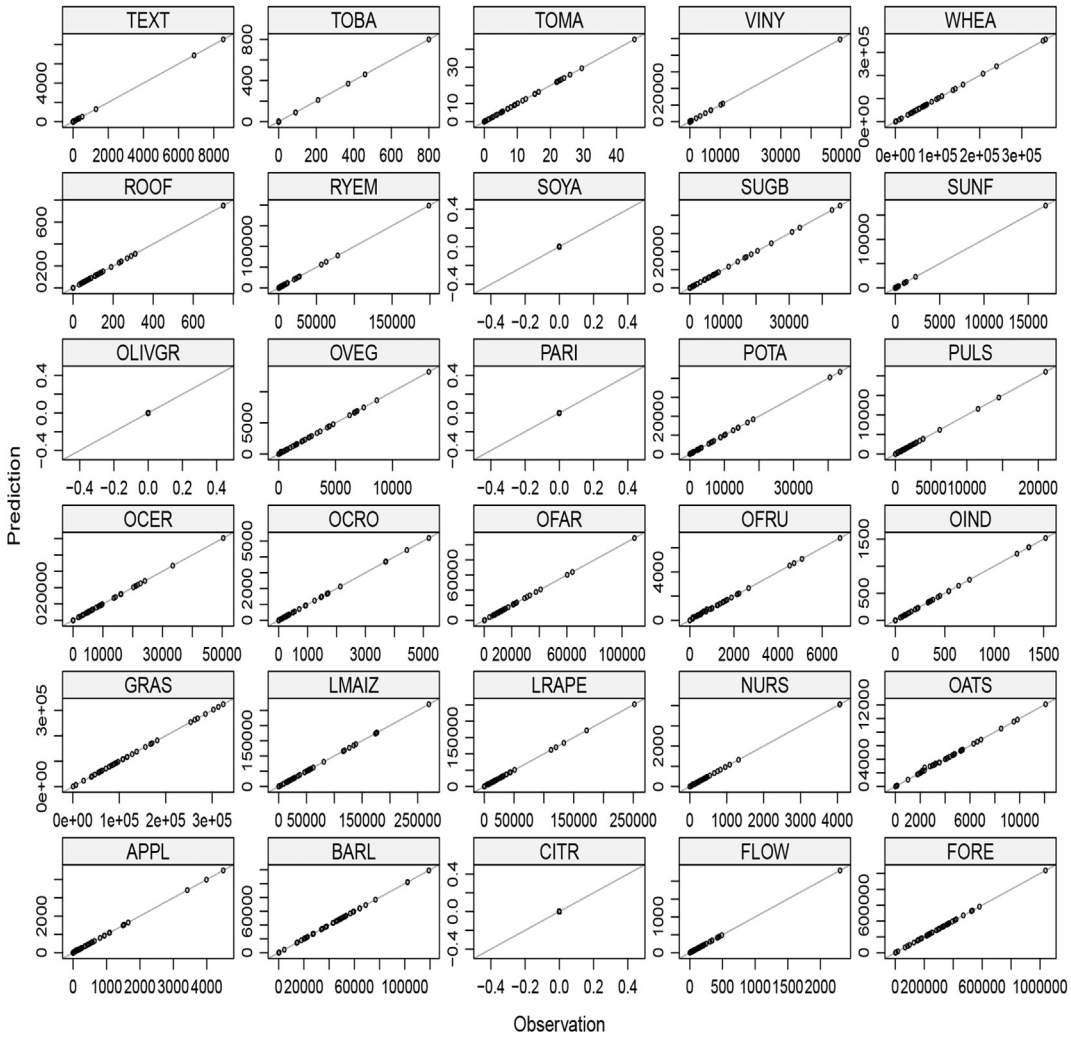

Fig. F.15. Q-Q plots of aggregated predictions of land-use areas (in ha) versus FSS observations at NUTS2 level for Germany.

Appendix G. Validation of the predictions at HSU level

Figs. G.16, G.17, G.18, G.19 provide all the summaries (minimum, first quartile, median, mean, third quartile and maximum) of the predictions errors at HSU level.

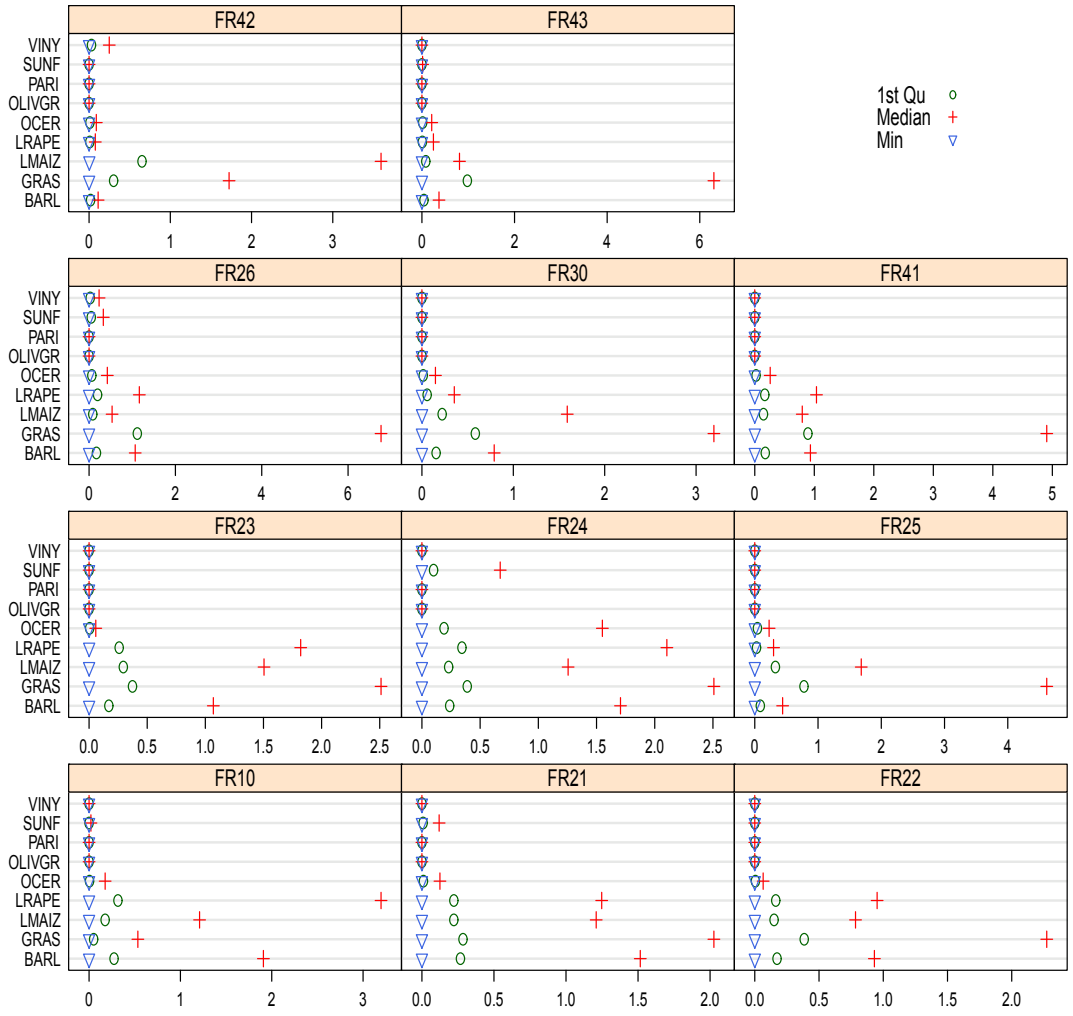

Fig. G.16. Summaries (minimum, first quartile and median) of the error terms (in $\times 100$ ha) of the predictions at HSU level using equation (6.15) for the first 11 NUTS2 regions in France (FR10, FR21, …, FR43). The land uses BARl, GRAS, LMAIZ, LRAPE, OCER, OLIVGR, PARI, SUNF and VINY stand respectively for barley, grassland, maize, rapeseed, other cereal, olive, rice, sunflowers and vineyard.

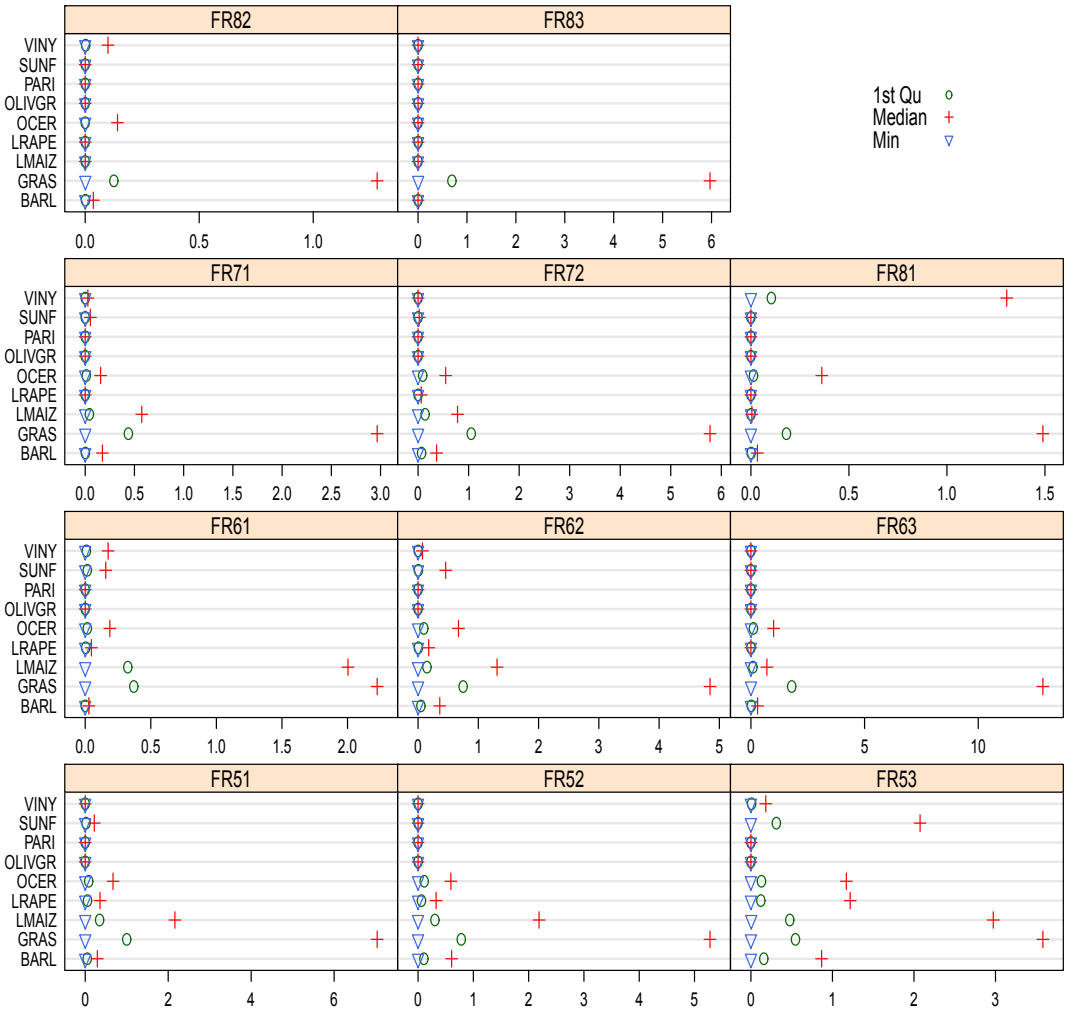

Fig. G.17. Summaries (minimum, first quartile and median) of the error terms (in $\times 100$ ha) of the predictions at HSU level using equation (6.15) for the last 11 NUTS2 regions in France (FR51, FR52, …, FR83). The land-uses BARI, GRAS, LMAIZ, LRAPE, OCER, OLIVGR, PARI, SUNF and VINY stand respectively for barley, grassland, maize, rapeseed, other cereal, olive, rice, sunflowers and vineyard.

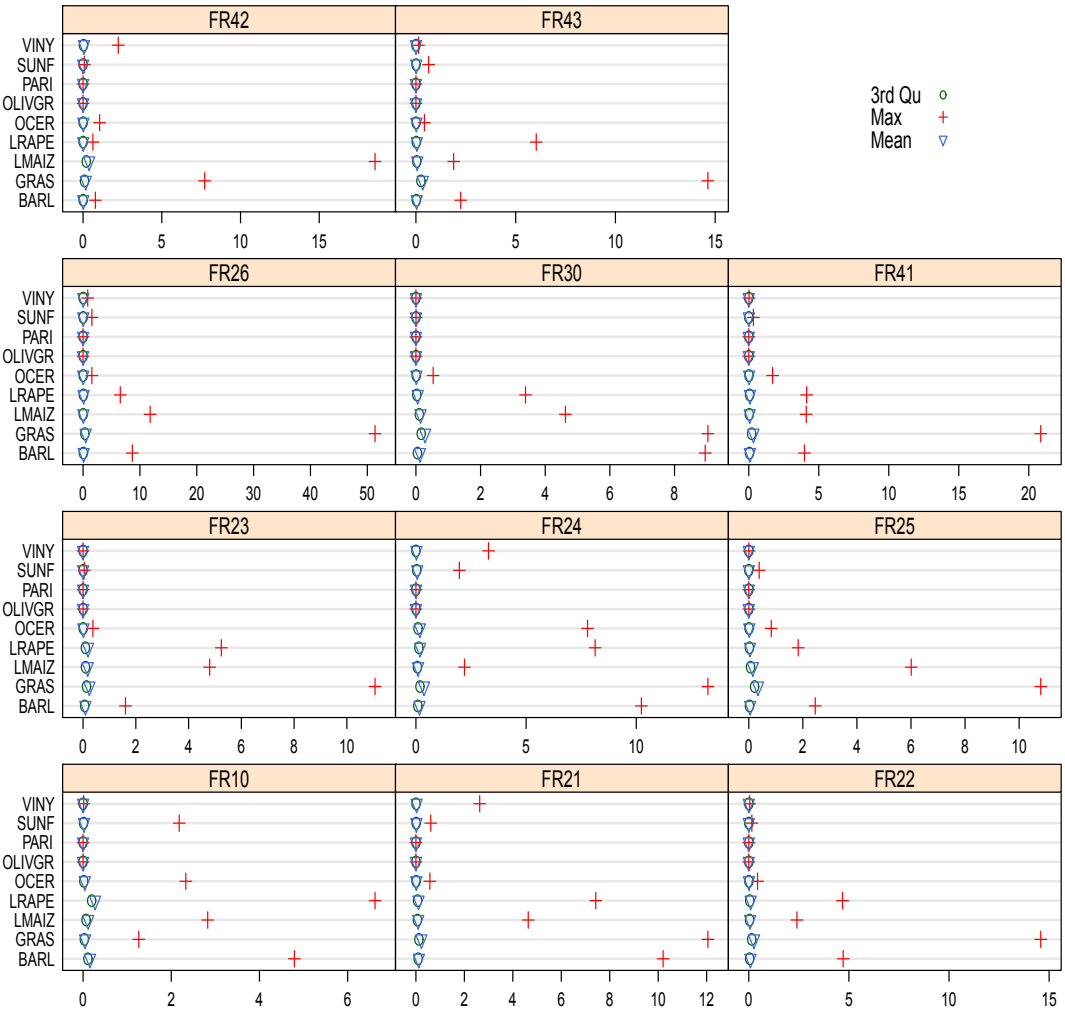

Fig. G.18. Summaries (mean, third quartile and maximum) of the error terms (in $\times 100$ ha) of the predictions at HSU level using equation (6.15) for the first 11 NUTS2 regions in France (FR10, FR21, …, FR43). The land-uses BARL, GRAS, LMAIZ, LRAPE, OCER, OLIVGR, PARI, SUNF and VINY stand respectively for barley, grassland, maize, rapeseed, other cereal, olive, rice, sunflowers and vineyard.

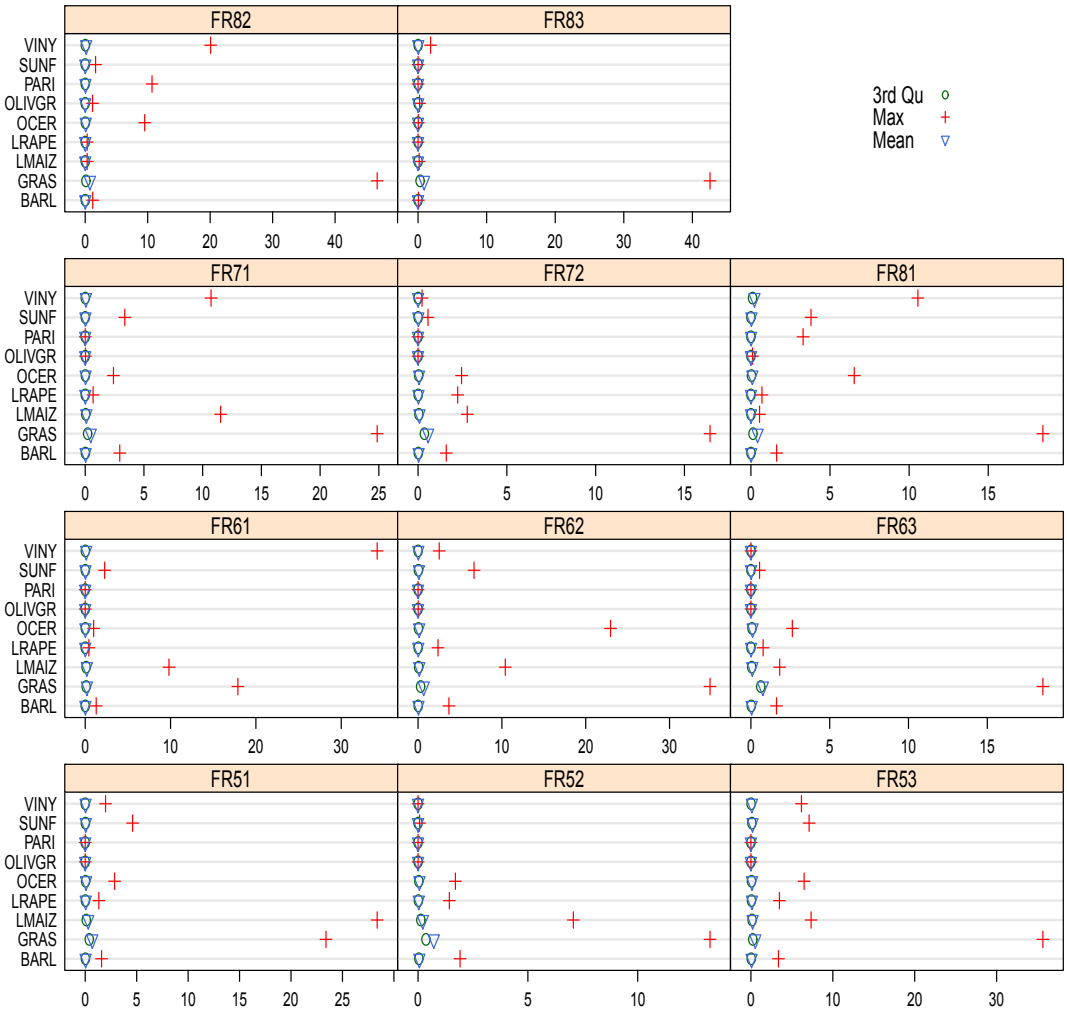

Fig. G.19. Summaries (mean, third quartile and maximum) of the error terms (in $\times 100$ ha) of the predictions at HSU level using equation (6.15) for the last 11 NUTS2 regions in France (FR51, FR52, …, FR83). The land-uses BARl, GRAS, LMAIZ, LRAPE, OCER, OLIVGR, PARI, SUNF and VINY stand respectively for barley, grassland, maize, rapeseed, other cereal, olive, rice, sunflowers and vineyard.

Appendix H. Predictions Uncertainties at HSU level

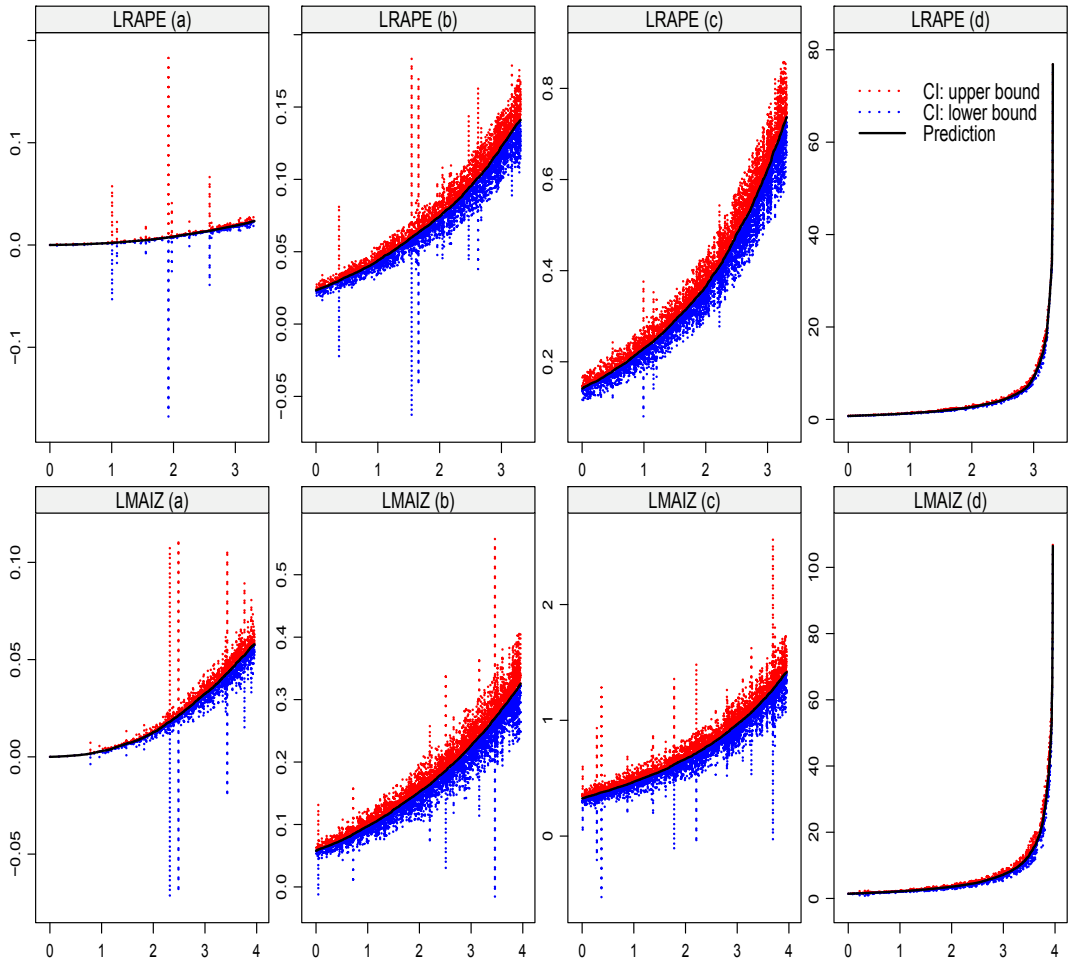

Fig. H.20. Uncertainties (95 % confident intervals in $\times 100$ ha) of predictions at HSU level after sorting and dividing all the HSUs into four groups according to the predicted areas of a given land-use and their quartiles. The land-uses LRAPE and LMAIZ stand respectively for rapeseed and maize (see paragraph 6.3). The *x*-coordinate reports the number of HSUs divided by 1000.

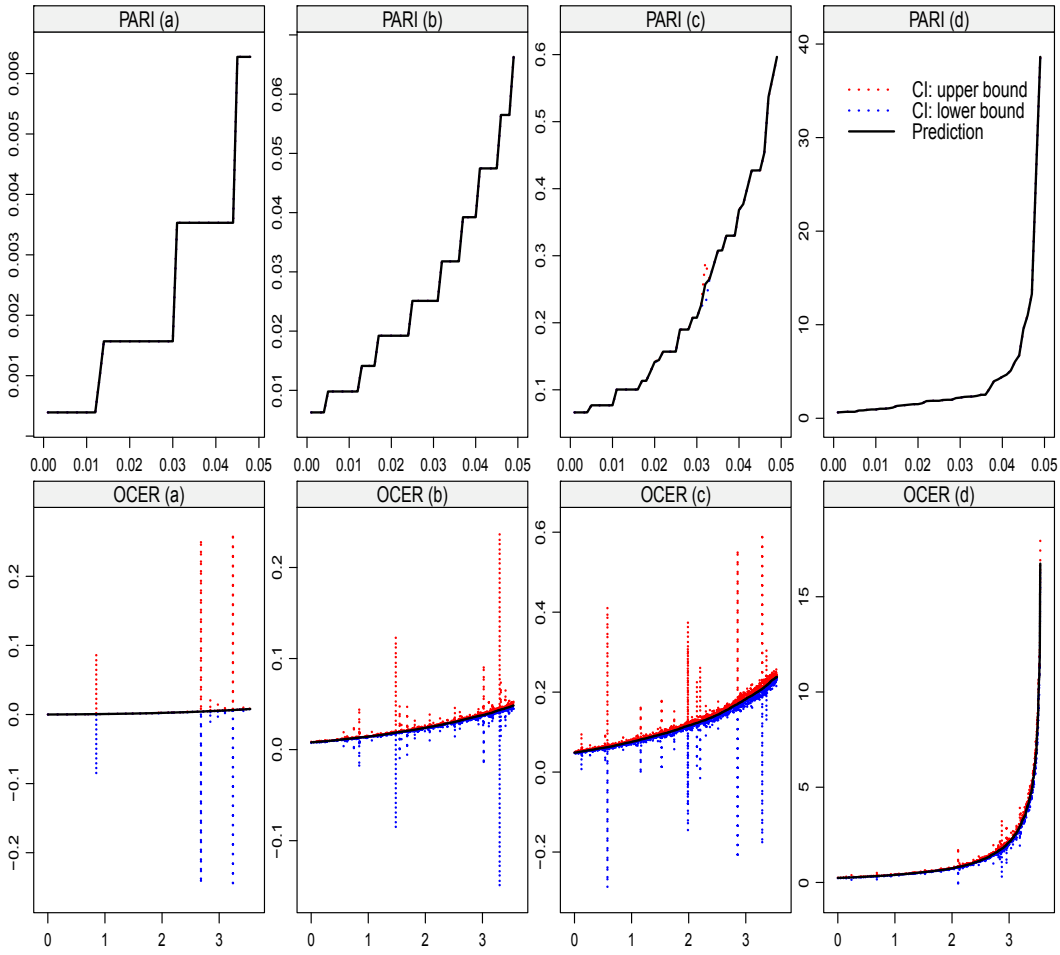

Fig. H.21. Uncertainties (95 % confident intervals in ×100 ha) of predictions at HSU level after sorting and dividing all the HSUs into four groups according to the predicted areas of a given land-use and their quartiles. The land-uses OCER and PARI stand respectively for other cereals and rice (see paragraph 6.3). The *x*-coordinate reports the number of HSUs divided by 1000.

Appendix I. Sensitivity indices

Figs. I.22, I.23, I.24, I.25 show the sensitivity indices of the parameters for the 22 models (one per NUTS2 region). We show the indices above 0.0005.

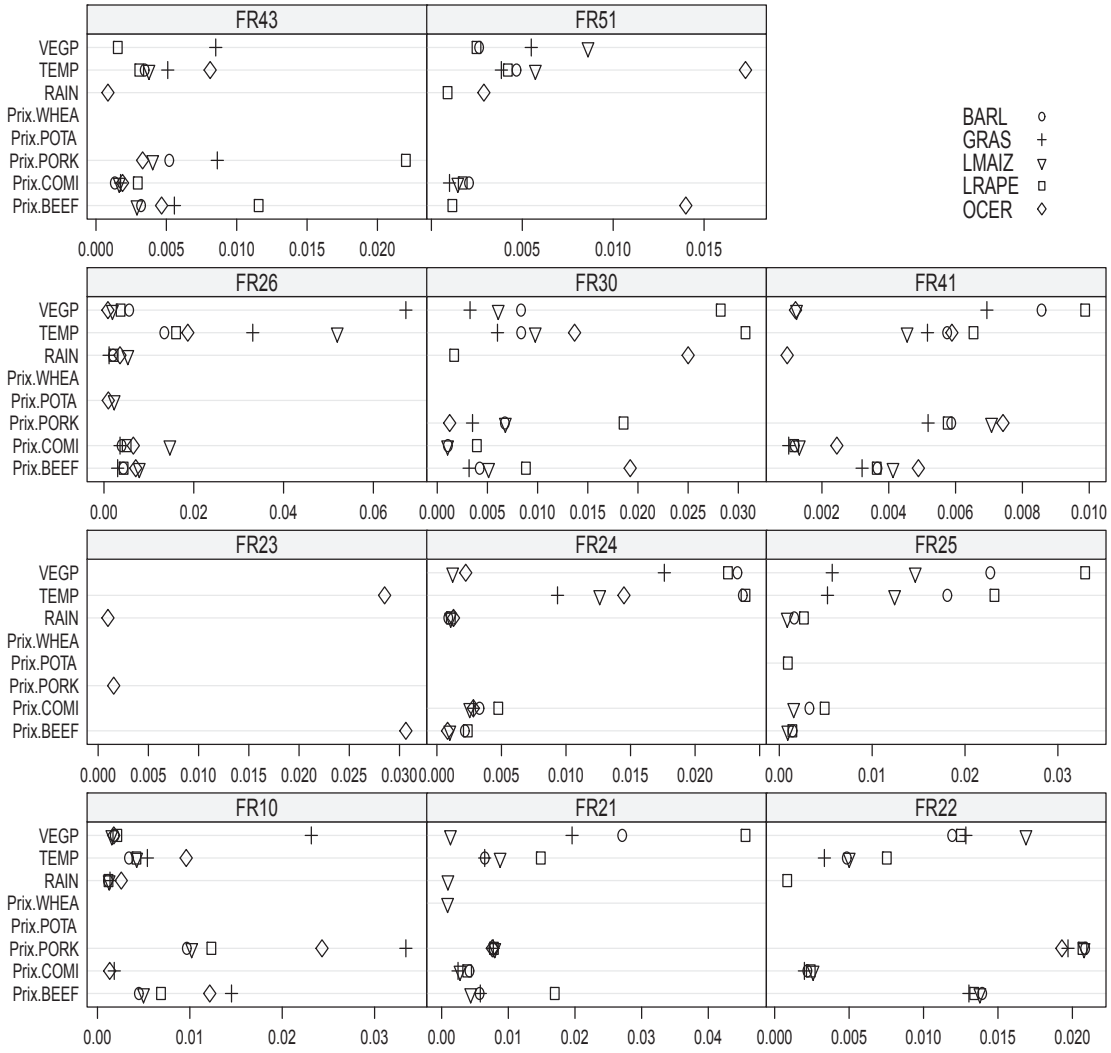

Fig. I.22. Sensitivity indices of parameters associated to explanatory variables for the first NUTS2 regions. The land-uses BARL, GRAS, LMAIZ, LRAPE, OCER stand respectively for barley, grassland, maize, rapeseed, other cereals (see paragraph 6.3).

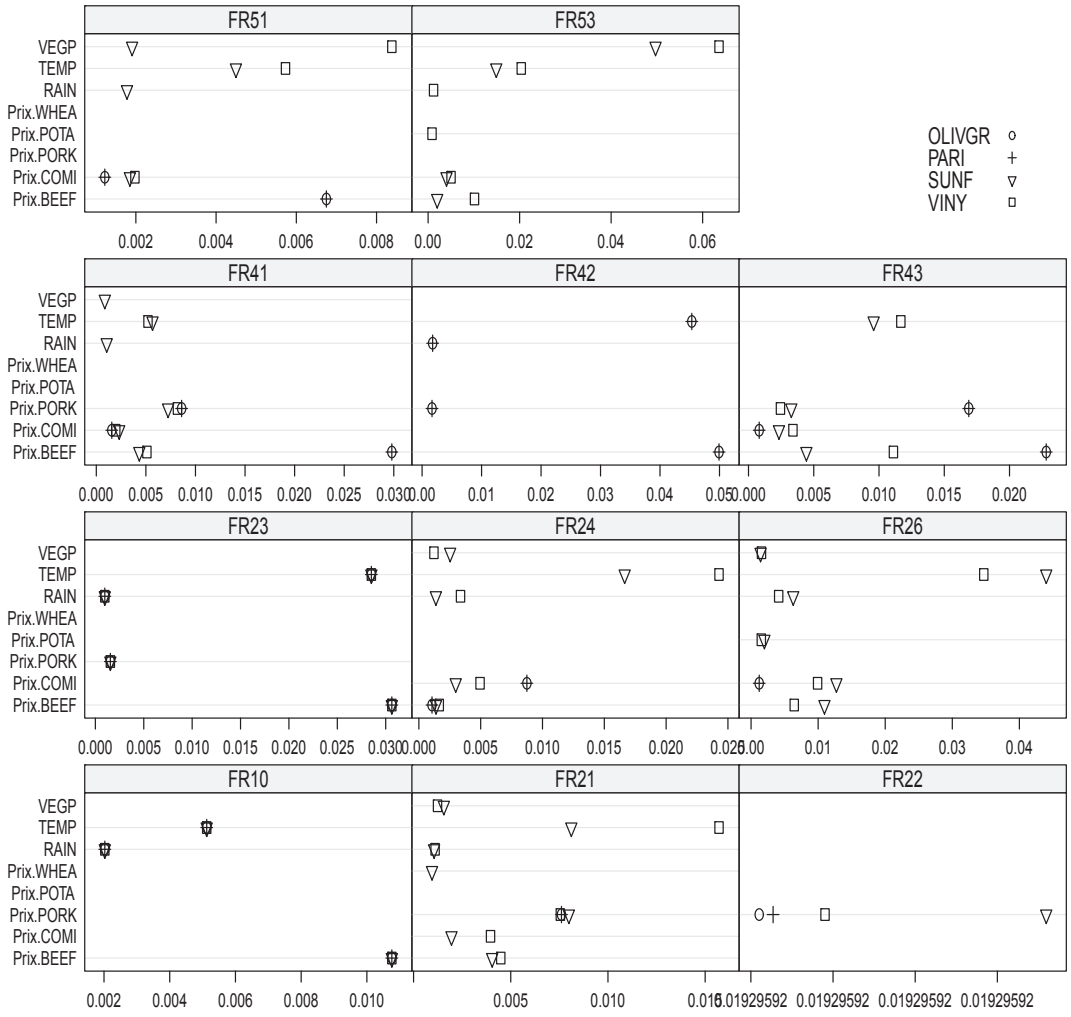

Fig. I.23. Sensitivity indices of parameters associated to explanatory variables for the last NUTS2 regions. The land-uses BARL, GRAS, LMAIZ, LRAPE, OCER stand respectively for barley, grassland, maize, rapeseed, other cereals (see paragraph 6.3).

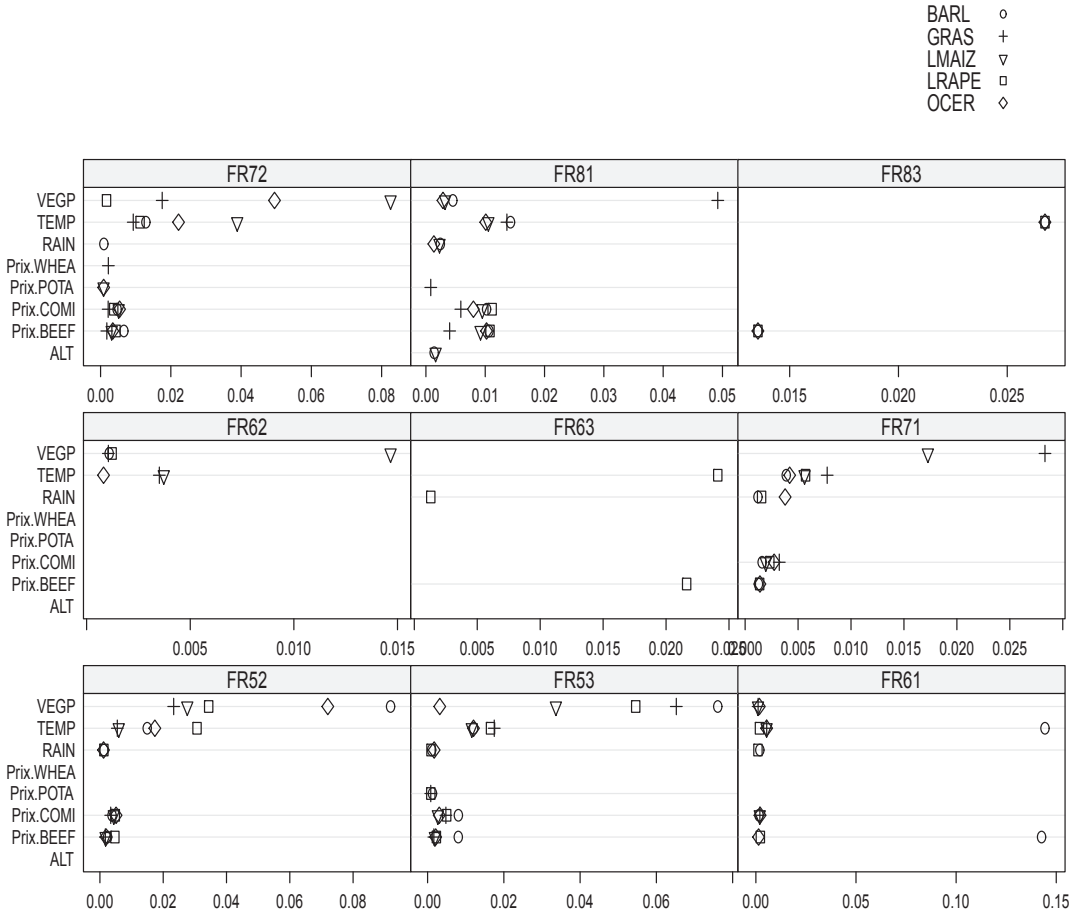

Fig. I.24. Sensitivity indices of parameters associated to explanatory variables for the first NUTS2 regions. The land-uses OLIVGR, PARI, SUNF and VINY stand respectively for olive, rice, sunflowers and vineyard (see paragraph 6.3).

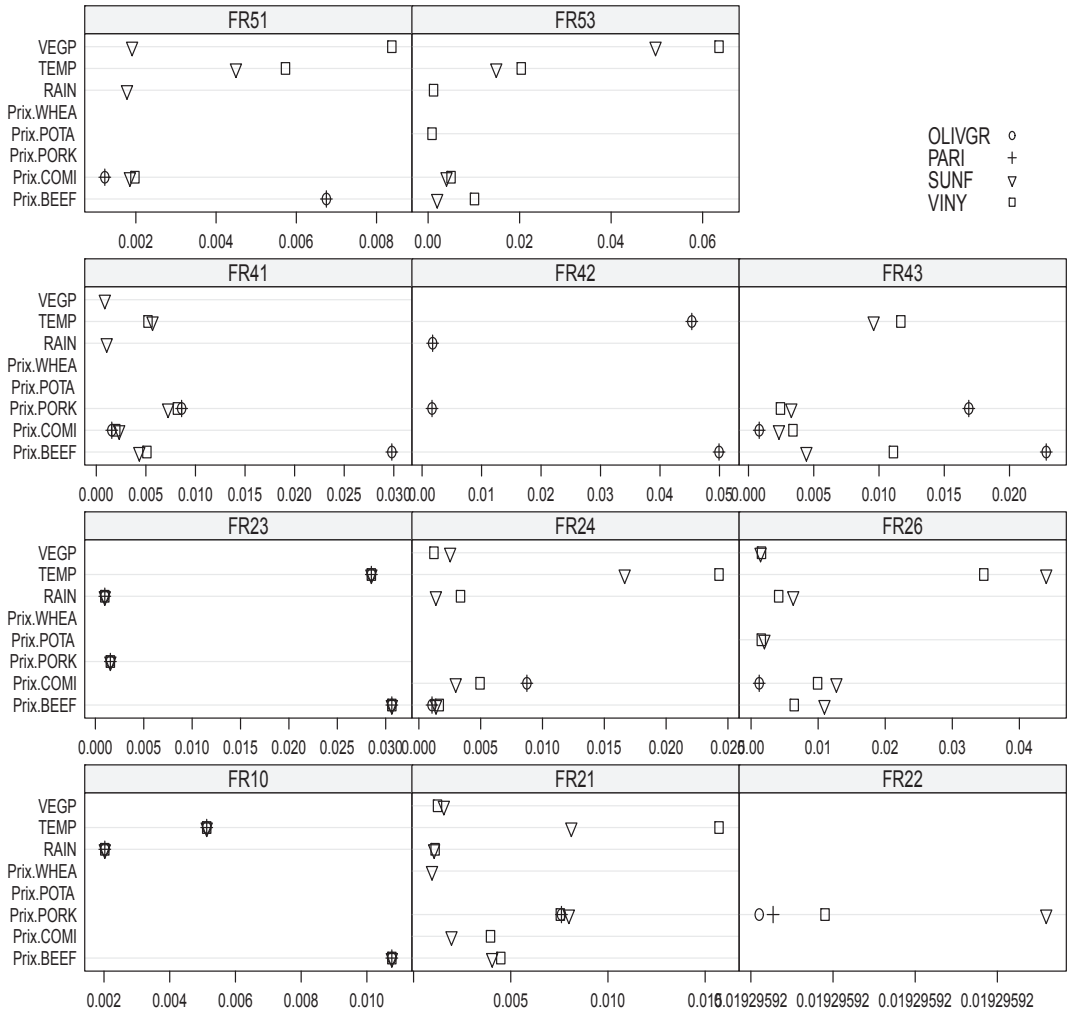

Fig. I.25. Sensitivity indices of parameters associated to explanatory variables for the last NUTS2 regions. The land-uses OLIVGR, PARI, SUNF and VINY stand respectively for olive, rice, sunflowers and vineyard (see paragraph 6.3).

Appendix J. Re-classification of Land-use classes from LUCAS, 2009 survey to match with CAPRI land-uses

Table J.4
Land-use/cover codes and their descriptions from LUCAS data. CAPRI names are the re-classification of the land-use/cover.

Land-use codes	Description	CAPRI names
A11	Buildings with 1–3 floors	OTHE
A12	Buildings with more than 3 floors	OTHE
A13	Greenhouses	OTHE
A21	Non Built up area features	OTHE
A22	Non built up linear features	OTHE
B11	Common wheat	SWHE
B12	Durum wheat	DWHE
B13	Barley	BARL
B14	Rye	RYEM
B15	Oats	OATS
B16	Maize	LMAIZ
B17	Rice	PARI
B18	Triticale	OCER
B19	Other cereals	OCER
B21	Potatoes	POTA
B22	Sugar beet	SUGB
B23	Other root crops	ROOF
B31	Sunflower	SUNF
B32	Rape and turnip seeds	LRAPE
B33	Soya	SOYA
B34	Cotton	TEXT
B35	Other fibre and oleaginous crops	TEXT
B36	Tobacco	TOBA
B37	Other non permanent industrial crops	OIND
B41	Dry pulses	PULS
B42	Tomatoes	TOMA
B43	Other fresh vegetables	OVEG
B44	Floriculture and ornamental plants	FLOW
B45	Strawberries	OVEG
B51	Clovers	OFAR
B52	Lucerne	OFAR
B53	Other legumes and mixtures for fodder	OFAR
B54	Mixed cereals for fodder	OFAR
B55	Temporary grasslands	OFAR
B71	Apple fruit	APPL
B72	Pear fruit	APPL
B73	Cherry fruit	APPL
B74	Nuts trees	OFRU
B75	Other fruit trees and berries	OFRU
B76	Oranges	CITR
B77	Other citrus fruit	CITR
B81	Olive groves	OLIVGR
B82	Vineyards	VINY
B83	Nurseries	NURS
B84	Permanent industrial crops	OCRO
BX1	Arable land (only in case PI)	OCRO
BX2	Permanent crops (only in case PI)	OCRO
C10	Broadleaved forest	FORE
C10	Broadleaved forest	FORE
C10	Broadleaved forest	FORE
C20	Coniferous forest	FORE
C20	Coniferous forest	FORE
C30	Mixed forest	FORE
C30	Mixed forest	FORE
D10	Shrubland with sparse tree cover	FORE
D20	Shrubland without tree cover	FORE
E10	Grassland with sparse tree/shrub cover	GRAS
E20	Grassland without tree/shrub cover	GRAS
E30	Spontaneously vegetated surfaces	GRAS
F00	Bare land	OTHE
G10	Inland water bodies	OTHE
G20	Inland running water	OTHE
G30	Coastal water bodies	OTHE
G50	Glaciers, permanent snow	OTHE
H11	Inland marshes	OTHE
H12	Peat bogs	OTHE
H21	Salt marshes	OTHE
H22	Salines	OTHE
H23	Intertidal flats	OTHE

Appendix K. Harmonized CORINE classes.

Table K.5
Description of CORINE classes and the harmonized classes used in this paper.

CORINE description	CORINE classes
Continuous urban fabric	URBA
Discontinuous urban fabric	URBA
Industrial or commercial units	URBA
Road and rail networks and associated land	URBA
Port areas	URBA
Airports	URBA
Mineral extraction sites	URBA
Dump sites	URBA
Construction sites	URBA
Green urban areas	URBA
Sport and leisure facilities	URBA
Non-irrigated arable land	ARAB
Permanently irrigated land	ARAB
Rice fields	RICF
Vineyards	VINY
Fruit trees and berry plantations	FTBP
Olive groves	OLIG
Pastures	PAST
Annual crops associated with permanent crops	HETC
Complex cultivation patterns	HETC
Land occupied by agriculture, with significant areas of natural vegetation	HETC
Agro-forestry areas	HETC
Broad-leaved forest	FORE
Coniferous forest	FORE
Mixed forest	FORE
Natural grasslands	GRAS
Moors and heathland	SHRU
Sclerophyllous vegetation	SHRU
Transitional woodland-shrub	SHRU
Beaches, dunes, sands	OPEN
Bare rocks	OPEN
Sparsely vegetated areas	OPEN
Burnt areas	OPEN
Glaciers and perpetual snow	OPEN
Inland marshes	INLW
Peat bogs	INLW
Salt marshes	INLW
Salines	INLW
Intertidal flats	INLW
Water courses	WATER
Water bodies	WATER
Coastal lagoons	WATER
Estuaries	WATER
Sea and ocean	WATER

References

Annoni, A., 2005. Proposal for a european grid system. In: European Reference Grids Workshop Proceedings and Recommendations. Ispra, Italy. URL EUR-Report21494EN. http://www.ec-gis.org/sdi/publist/pdfs/annoni2005eurgrids.pdf.

Beven, K., Freer, J., 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the glue methodology. J. Hydrol. 249, 1–29.

Bordogna, G., Boschetti, M., Brivio, P., Carrara, P., Stroppiana, D., Weissteiner, C., 2012. Handling heterogeneous bipolar information for modelling environmental syndromes of global change. Environ. Model. Softw. 36, 131–147.

Britz, W., Leip, A., 2009. Development of marginal emission factors for n losses from agricultural soils with the dndc-capri meta-model. Agric. Ecosyst. Environ. 133 (3–4), 267–279.

Brodlie, K., Allendes Osorio, R., Lopes, A., 2012. A review of uncertainty in data visualization. In: Dill, J., Earnshaw, R., Kasik, D., Vince, J., Wong, P.C. (Eds.), Expanding the Frontiers of Visual Analytics and Visualization, pp. 81–109.

Butterbach-Bahl, K., Gundersen, P., Ambus, P., Augustin, J., Beier, C., Boeckx, P., Dannemann, M., Gimeno, B., Kiese, R., Kitzler, B., Ibrom, A., Rees, R., Smith, K., Stevens, C., Vesala, T., Zechmeister-Boltenstern, S., 2011. Nitrogen Processing in the Biosphere. Cambridge University Press, pp. 99–125.

Cantelaube, P., Carles, M., 2015. Le registre parcellaire graphique: des données géographiques pour décrire la couverture du sol agricole. In: Cahier des Techniques de l'INRA, Special issue GéoExpé, pp. 58–64.

Chakir, R., 2009. Spatial downscaling of agricultural land-use data: an econometric approach using cross entropy. Land Econ. 85, 238–251.

Cleveland, W.S., 1979. Robust locally weighted regression and smoothing scatterplots. J. Am. Stat. Assoc. 74 (368), 829–836.

Cleveland, W.S., Devlin, S.J., 1988. Locally weighted regression: an approach to regression analysis by local fitting. J. Am. Stat. Assoc. 83 (403), 596–610.

de Rigo, D., 2013. Software Uncertainty in Integrated Environmental modelling: the Role of Semantics and Open Science. CoRR abs/1311.4762.

Dönmez, D., Grote, G., 2011. Managing uncertainty in software development projects: an assessment of the agile development method scrum. In: Sillitti, A., Hazzan, O., Bache, E., Albaladejo, X. (Eds.), Agile Processes in Software Engineering and Extreme Programming, Vol. 77 of Lecture Notes in Business Information Processing, pp. 326–328.

Dubois, D., Prade, H., Esteva, F., Garcia, P., Godo, L., 1997. A logical approach to interpolation based on similarity relations. Int. J. Approx. Reason. 17 (1), 1–36.

EC-JRC-AGRI4CAST, 2012. Interpolated Meteorological Data. European Commission Joint Research Centre, Institute for Environment and Sustainability, Monitoring Agricultural Resources (MARS) Unit. URL. http://mars.jrc.ec.europa.eu/mars/.

EEA, 2012. Annual European Community Greenhouse Gas Inventory 1990-2010 and Inventory Report 2012. Submission to the unfccc secretariat. European Environment Agency report, Kongens Nytorv 6, 1050 København K, Danemark.

ESA, 2008. European Space Agency: Globcover Land Cover Database, Project, Led by medias-france. URL. http://due.esrin.esa.int/page_globcover.php.

ETCSIA, 2012. European Topic Centre on Spatial Information and Analysis: Corine Land Cover 2006 Raster Data.

EUROSTAT, 2010. Farm Structure Survey (FSS). URL. http://epp.eurostat.ec.europa.eu/portal/page/portal/agriculture/farm_structure/database.

FAO/IIASA/ISRIC/ISS-CAS/JRC, 2009. Harmonized World Soil Database, Version 1.1. FAO, Rome, Italy and IIASA, Laxenburg, Austria.

Follador, M., Leip, A., Orlandini, L., 2011. Assessing the impact of cross compliance measures on nitrogen fluxes from european farmlands with dndc-europe. Environ. Pollut. 159 (11), 3233–3242. Assessment of Nitrogen Fluxes to Air and Water from Site Scale to Continental Scale.

Frühwirth-Schnatter, S., Frühwirth, R., 2012. Bayesian inference in the multinomial logit model. Austrian J. Stat. 41 (1), 27–43.

Gocht, A., Röder, N., 2014. Using a bayesian estimator to combine information from a cluster analysis and remote sensing data to estimate high-resolution data for agricultural production in germany. Int. J. Geogr. Inf. Sci. 28 (9), 1744–1764.

Goldfarb, D., Idnani, A., 1983. A numerically stable dual method for solving strictly convex quadratic programs. Math. Program. 27 (1), 1–33.

Haylock, M.R., Hofstra, N., Klein Tank, A.M.G., Klok, E.J., Jones, P.D., New, M., 2008. A European daily high-resolution gridded data set of surface temperature and precipitation for 1950-2006. J. Geophys. Res. Atmos. 113 (D20).

Hofstra, N., Haylock, M., New, M., Jones, P.D., 2009. Testing E-OBS European high-resolution gridded data set of daily precipitation and surface temperature. J. Geophys. Res. Atmos. 114 (D21).

Hosmer, D., Lemeshow, S., 2000. Applied Logistic Regression, second ed. Wiley.

INSPIRE, 2008. Reference systems thematic working group: D2.8.i.2 specifications on geographical grid systems - draft guidelines. In: Workshop Proceedings. URL. http://inspire.ec.europa.eu/reports/ImplementingRules/DataSpecifications/INSPIRE_Specification_GGS_v2.0.pdf.

Jarvis, A., Reuter, H.-I., Nelson, A., Guevara, E., 2008. Hole-filled Srtm for the Globe Version 4. Available from the CGIAR-CSI SRTM 90 m Database. URL. http://srtm.csi.cgiar.org.

Kempen, M., Heckelei, T., Britz, W., 2005. An econometric approach for spatial disaggregation of crop production in the EU. In: Working Paper Presented at the EAAE Seminar. University of Bonn, Institute for Agricultural Policy, Market Research and Economic Sociology, Parma, 3–5 February 2005.

Kempeneers, P., McInerney, D., Sedano, F., Gallego, J., Strobl, P., Kay, S., Korhonen, K., San-Miguel-Ayanz, J., 2013. Accuracy assessment of a remote sensing-based, pan-european forest cover map using multi-country national forest inventory data. IEEE J. Sel. Top. Appl. Earth Observations Remote Sens. 6 (1), 54–65.

Kotz, S., Balakrishnan, N., Johnson, N.L., 2005. Multivariate Logistic Distributions. John Wiley & Sons, Inc., pp. 551–576

Lamboni, M., 2013. Bayesian prediction of crop land shares for environmental impact assessment: case of Norway. In: ICAS VI - Sixth International Conference on Agricultural Statistics, Rio de Janeiro, Brazil.

Lamboni, M., Koeble, R., Leip, A., 2013. Prediction of crop land shares for environmental impact assessment over EU-27. In: ICAS VI - Sixth International Conference on Agricultural Statistics, Rio de Janeiro, Brazil.

Lamboni, M., Koeble, R., Leip, A., 2014a. Bayesian spatial disaggregating of shares: application to land use shares in EU. In: 29th International Workshop on Statistical Modelling, Göttingen, Germany.

Lamboni, M., Makowski, D., Lehuger, S., Gabrielle, B., Monod, H., 2009. Multivariate global sensitivity analysis for dynamic crop models. Field Crops Res. 113, 312–320.

Lamboni, M., Makowski, D., Monod, H., 2008. Multivariate Global Sensitivity Analysis for Discrete-time Models. Rapport technique 2008-3, INRA, UR341. Mathématiques et Informatique Appliquées, Jouy-en-Josas, France.

Lamboni, M., Makowski, D., Monod, H., 2011a. Indices de sensibilité et sélection de paramètres: des liaisons dangereuses pour l'erreur quadratique de prédiction? J. de la Société Française De Statistique 152, 47–69.

Lamboni, M., Monod, H., Makowski, D., 2011b. Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models. Reliab. Eng.

Syst. Saf. 96, 450–459.
Lamboni, M., Sanaa, M., Tenenhaus-Aziza, F., 2014b. Sensitivity analysis for critical control points determination and uncertainty analysis to link fso and process criteria: application to listeria monocytogenes in soft cheese made from pasteurized milk. Risk Anal. 34 (4), 751–764.
Lehman, M.M., 2000. Rules and Tools for Software Evolution and Management. In: Pre-prints of FEAST 2000 International Workshop on Feedback in Software and Business Processes. Imperial College, London.
Lehman, M.M., Belady, L., 1985. Software Evolution - Processes of Software Change. Academic Press, London, Imperial College, London.
Lehuger, S., Gabrielle, B., Laville, P., Lamboni, M., Loubet, B., Cellier, P., 2011. Predicting and mitigating the net greenhouse gas emissions of crop rotations in western Europe. Agric. For. Meteorol. 151 (12), 1654–1671.
Leip, A., 2011. Assessing the Environmental Impact of Agriculture in Europe: the Indicator Database for European Agriculture. ASC, pp. 371–385. Ch. 19.
Leip, A., Britz, W., Bulgheroni, C., Carmona-Garcia, G., Koeble, R., Lamboni, M., Paracchini, M., Ramos, F., Saban-Ozbek, F., Terres, J.-M., Wania, A., Weiss, F., 2013. Capri - a spatial assessment tool for agri-environmental indicators in the EU. In: ICAS VI - Sixth International Conference on Agricultural Statistics, Rio de Janeiro, Brazil.
Leip, A., Marchi, G., Koeble, R., Kempen, M., Britz, W., Li, C., Jan 2008. Linking an economic model for european agriculture with a mechanistic model to estimate nitrogen and carbon losses from arable soils in Europe. Biogeosciences 5 (1), 73–94.
Li, B.-L., 2000. Fractal geometry applications in description and analysis of patch patterns and patch dynamics. Ecol. Model. 132 (1–2), 33–50.
Lindley, D.V., Smith, A.F.M., 1972. Bayes estimates for the linear model. J. R. Stat. Soc. Ser. B Methodol. 34 (1), 1–41.
Mandelbrot, B., 1983. The Fractal Geometry of Nature/Revised and Enlarged Edition. Freeman and Co, New York, W.H.
Mas, J.-F., Pérez-Vega, A., Clarke, K.C., 2012. Assessing simulated land use/cover maps using similarity and fragmentation indices. Ecol. Complex. 11, 38–45.
McFadden, D., 1974. Conditional Logit Analysis of Qualitative Choice Behavior. Academic Press, New York, pp. 105–142.
O'Hagan, A., 2012. Probabilistic uncertainty specification: overview, elaboration techniques and their application to a mechanistic model of carbon flux. Environ. Model. Softw. 36, 35–48.
Pekkarinen, A., Reithmaier, L., Strobl, P., 2009. Pan-european forest/non-forest mapping with landsat etm+ and CORINE land cover 2000 data. ISPRS J. Photogrammetry Remote Sens. 64 (2), 171–183.
Powers, D., 2011. Evaluation: from precision, recall and f-factor to roc, informedness, markedness & correlation. J. Mach. Learn. Technol. 2, 37–63.
Pérez-Vega, A., Mas, J.-F., Ligmann-Zielinska, A., 2012. Comparing two approaches to land use/cover change modeling and their implications for the assessment of biodiversity loss in a deciduous tropical forest. Environ. Model. Softw. 29 (1), 11–23.
Röder, N., Gocht, A., 2013. Recovering localised information on agricultural structures while observing data confidentiality regulations - the potential of different data aggregation and segregation techniques. J. Land Use Sci. 8 (1), 31–46.
Reibel, M., Agrawal, A., 2007. Areal interpolation of population counts using pre-classified land cover data. Popul. Res. Policy Rev. 26 (5–6), 619–633.
Rinderknecht, S.L., Borsuk, M.E., Reichert, P., 2012. Bridging uncertain and ambiguous knowledge with imprecise probabilities. Environ. Model. Softw. 36, 122–130.
Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S., 2008. Global Sensitivity Analysis: the Primer. Wiley.
Sellers, P.J., Heiser, M.D., Hall, F.G., Verma, S.B., Desjardins, R.L., Schuepp, P.M., MacPherson, J.I., 1997. The impact of using area-averaged land surface properties -topography, vegetation condition, soil wetness-in calculations of intermediate scale (approximately 10 km2) surface-atmosphere heat and moisture fluxes. J. Hydrol. 190 (3–4), 269–301.
Smith, A.F.M., 1973. A general bayesian linear model. J. R. Stat. Soc. Ser. B Methodol. 35 (1), 67–75.
Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions (with discussion). J. R. Stat. Soc. B 36, 111–147.
Thomas-Agnan, C., Vanhemsz, A., 2013. Spatial Reallocation of Areal Data – a Review. Working Papers. Toulouse School of Economics.
Yang, Y., 2007. Consistency of cross validation for comparing regression procedures. Ann. Stat. 35, 2450–2473.